Annotation of Conceptual Co-reference and Text Mining the Qur'an

Abdul Baquee Muhammad

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Computing

September, 2012

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Dedication

أهدي هذه الأطروحة للدكتور سفر بن عبد الرحمن الحوالي. رئيس قسم العقيدة بجامعة أم القرى سابقا. لقد ظلت فكرة الالتحاق ببرنامج الدكتوراه حلماً يدور في خاطري إلى أن التقيت به. فكان لدعمه وتشجيعه الأثر الكبير لإتمام هذا البحث.

I dedicate this thesis to Dr. Safar ibn Abdur-Rahman Al-Hawali – former dean of Islamic theology Department, Umm Al-Qura University, Makkah, Saudi Arabia. Pursuing a PhD journey was a dream until I met Dr. Safar. His support and encouragement made this happen.

Declaration

I declare that the work presented in this thesis, is the best of my knowledge of the domain, original, and my own work. Most of the work presented in this thesis have been published. When co-authored with my supervisor, I prepared the first draft and then Eric reviewed and suggested amendments. Publications are listed below:

(Abdul Baquee Muhammad[1])

Chapter 2 Sections 2.1 and 2.2

Muhammad, A. and Atwell, E. (2009) A Corpus-based computational model for knowledge representation of the Qur'an. 5th Corpus Linguistics Conference, Liverpool.

Muhammad, A. and Atwell, Eric, (2012) "QurAna: corpus of the Qur'an annotated with pronominal anaphora", LREC 2012

Atwell, ES; Brierley, C; Dukes, K; Sawalha, M; **Muhammad, A** (2011) An Artificial Intelligence Approach to Arabic and Islamic Content on the Internet in: Proceedings of NITS 3rd National Information Technology Symposium. 2011.

Chapter 4 Section 4.3

Muhammad, A. and Atwell, E. (2009) A Corpus-based computational model for knowledge representation of the Qur'an. 5th Corpus Linguistics Conference, Liverpool.

Chapter 5

Muhammad, A. and Atwell, Eric, (2012) "QurAna: corpus of the Qur'an annotated with pronominal anaphora", LREC 2012

Chapter 6

Muhammad, A. and Atwell, Eric. (2012) "QurSim: A corpus for evaluation of relatedness in short texts", LREC 2012

Chapter 8

---

[1] My surname was officially changed at the University records from 'Sharaf' to 'Muhammad' to resolve earlier confusion between surname and father's name in my passport. Although, in this section I have used the corrected surname, readers may note that my earlier surname i.e.,'Sharaf' was used in most of the already published papers.

Muhammad, A and Atwell, Eric (2011) التصنيف الآلي للسور القرآنية "Automatic categorization of the Qur’anic chapters". 7th International Computing Conference in Arabic (ICCA11).31th May - 2nd June 2011, Imam Mohammed Ibn Saud University, Riyadh, Saudi Arabia

Chapter 9

Muhammad, A. and Atwell, E. (2009) A Corpus-based computational model for knowledge representation of the Qur'an. 5th Corpus Linguistics Conference, Liverpool.

In addition all text mining applications cited in chapter 7 were scripted by me and made available online along with their documentation under the website which I own:

http://textminingtheQuran.com/wiki

Acknowledgements

First and foremost I praise Allah for His bounty and blessings who provided me the strength to research His Words (i.e., the Qur'an) leveraging on the advancement of computational sciences.

I am indebted to my advisor, Dr. Eric Atwell, for the guidance he provided me during my studies. Eric has given me plenty of freedom to explore the directions I was most interested in. At the same time he brought me back to the right focused course, when temptations used to drag me to distant directions. Eric's comments on draft papers were very instrumental and resulted in quality publications. Eric also was very understanding during the ups and downs of my PhD journey, and I always found in his counselling and support a moral boost towards destination.

I would like to acknowledge Kais Dukes for his Qur'anic Arabic Corpus (QAC). I have used it extensively while annotating Qur'anic anaphora, and it has substantially reduced the manual effort I would had to exercise otherwise . I have enjoyed being part of the Natural Language Processing group at Leeds University. The weekly seminars were very informative and we used to exchange ideas and suggestions on each other's research. Special thanks to Majdi Sawalha and Justin Washtell for their constructive suggestions and support on various topics.

I would like to thank also the Islamic Development Bank – my employer- for granting me leave to study. My colleagues in the IT Department were very cooperative and supportive during the course of my studies.

This achievement was not possible without the support and understanding of all members of my family. My parents – Dr. Sharful Islam and Fatima Islam- were always behind me with their supplications and encouragement. Kids – Anas, Muhammad, Ayesha and Safiyyah- were very understanding and cooperative when dad spent yet another weekend behind the screen annotating or experimenting with datasets. And of course, my wife – Momtaj Begum- for her patience and support taking care of the kids and guaranteeing a conducive atmosphere for research at home.

Abstract

This research contributes to the area of corpus annotation and text mining by developing novel domain specific language resources. Most practical text mining applications restrict their domain. This research restricts the domain to the Qur'anic Text.

In this thesis, a number of pre-processing steps were undertaken and annotation information were added to the Qur'an. The raw Arabic Qur'an was pre-processed into morphological units using the Qur'anic Arabic Corpus (QAC). Qur'anic terms were indexed and converted into a vector space model using techniques in Information Retrieval (IR). In parallel, nearly 24,000 Qur'anic personal pronouns were annotated with information on their referents. These referents are consolidated and organized into a total of over 1,000 ontological concepts. Moreover, a dataset of nearly 8,000 pairs of related Qur'anic verses are compiled from books of scholarly commentary on the Qur'an. This vector space model, the pronoun tagging, the verse relatedness dataset, and the part-of-speech tags available in QAC all together served for a number of Qur'anic text mining applications which were rendered online for public use. Among these applications: lemma concordance, collocation, POS search of the Qur'an, verse similarity measures, concept clouds of a given verse, pronominal anaphora and Qur'anic chapter similarity.

Furthermore, machine learning experiments were conducted on automatic detection of verse similarity/relatedness as well as categorization of Qur'anic chapters based on their chronology of revelation. Domain specific linguistic features were investigated to induct learning algorithms. Results show that deep linguistic and world knowledge is needed to reach the human upper bound in certain computational tasks such as detecting text relatedness, question answering and textual entailment. However, many useful queries can be addressed using text mining techniques and layers of annotations made available through this research. The works presented here can be extended to include other similar texts like Hadith (i.e., saying of Prophet Muhammad), or other scriptures like the Gospels.

Table of Contents

Dedication ... **ii**
Declaration ... **iii**
Acknowledgements ... **v**
Abstract ... **vi**
Table of Contents ... **vii**
List of Tables ... **xiii**
List of Qur'anic Verses ... **xiv**
List of Figures ... **xvii**
Chapter 1 Introduction ... **1**
1.1 Proposed Text Mining Solution ... 1
1.2 Novel Contributions of this thesis ... 2
1.3 Potential Usage of the resources produced by this thesis ... 3
1.3.1 Arabic NLP ... 3
1.3.2 Computational Text similarity and relatedness ... 4
1.3.3 Computational Anaphora Resolution ... 4
1.3.4 Computational Stylistics ... 4
1.3.5 Translation studies ... 5
1.4 Thesis Outline ... 6
1.5 Summary ... 7
Chapter 2 Background on the language of the Qur'an ... **9**
2.1 What is the Qur'an? ... 9
2.1.1 Classical Arabic (CA) vs. Modern Standard Arabic (MSA) ... 9
2.1.2 Inimitability of the Qur'an ... 10
2.1.3 Some Linguistic features ... 10
2.1.3.1 Scattered information on the same topic ... 11
2.1.3.2 Literal vs. technical sense of a word ... 11
2.1.3.3 Verbs preposition binding ... 12
2.1.3.4 Metaphor and Figurative use ... 13
2.1.3.5 Metonymy ... 13
2.1.3.6 Imperative vs. non-Imperative senses ... 14
2.1.4 Why computational analysis of the Qur'an? ... 14
2.2 Pronominal Anaphora ... 17

2.2.1 Pronominal System in Arabic 17
2.2.2 Anaphora in the Qur'an 18
2.2.2.1 Pronoun antecedent agreement in the Qur'an..... 20
2.3 Text Similarity and relatedness in the Qur'an 22
2.3.1 Ambiguity in resolving the sense of a noun 22
2.3.2 Ambiguity in resolving the sense of a verb 23
2.3.3 Ambiguity in resolving the sense of a particle..... 24
2.3.4 Clarifying the meaning of an indefinite word..... 24
2.3.5 Clarifying an indefinite subject through a relative clause.... 25
2.3.6 Overriding Linguistic Defaults..... 25
2.3.7 Cause of an action in different Verse 26
2.3.8 Reason of an action in different Verse 26
2.3.9 Mentioning the Object of a Subject in different verse 27
2.3.10 Mentioning the Adverb of Place and Time in another verse 27
2.3.11 Specify Semantic Role in a different verse 28
2.3.12 Clarify a difficult vocabulary in another verse 28
2.3.13 Specifying whether a condition is being fulfilled 29
2.3.14 Explicit reference to another verse 29
2.4 The Qur'an Translations..... 30
2.5 Qur'an Exegesis (Tafsir books) 30
2.5.1 Significance of books of *Tafsir*..... 30
2.5.2 Tafsir Methodologies 32
2.6 Summary..... 33
Chapter 3 Literature Review 34
3.1 Computational Qur'anic Studies 34
3.1.1 Computational Clustering 34
3.1.1.1 Thabet 2005 34
3.1.1.2 Moisl 2009 35
3.1.2 Computational Stylistic Studies 37
3.2 Anaphora Resolution..... 38
3.2.1 Automatic resolution systems..... 39
3.3 Computational text similarity and relatedness 40
3.3.1 Evaluation Corpora for Text Similarity and Relatedness experiments..... 40

3.3.2 Knowledge Sources for similarity and relatedness experiments 41
3.3.3 Semantic relatedness measures 41
3.3.3.1 Path-Based Measures 41
3.3.3.2 Information Content based measures 42
3.3.3.3 Gloss based measure 42
3.3.3.4 Concept vector based measures 42
3.3.4 Text similarity and relatedness applications 42
3.4 Summary 43
Chapter 4 Linguistic Annotations of the Qur'an 44
4.1 Raw Qur'anic Text 44
4.1.1 Qur'anic Pause Marks 44
4.1.1.1 Mandatory Pause (M) 45
4.1.1.2 Voluntary Pause (V) 45
4.1.1.3 Voluntary Pause where continuation preferred (VC) 46
4.1.1.4 Voluntary Pause where pause preferred (VP) 46
4.1.1.5 Exclusive OR pause (XOR) 47
4.1.2 Qur'anic Pause Mark Annotation 47
4.2 Morphological Annotation 47
4.2.1 Haifa Corpus 47
4.2.1.1 Lexicon 48
4.2.1.2 Finite-State Rules 48
4.2.1.3 Query Interface 49
4.2.1.4 Evaluation 49
4.2.2 Qur'anic Arabic Corpus (QAC) 50
4.3 Syntactic Annotation 51
4.3.1 Qur'anic Arabic Dependency Treebank 51
4.3.2 Pronoun antecedent annotation 53
4.4 Semantic Annotation 54
4.4.1 Frame Semantics 54
4.4.2 Qur'anic Prepositional Verbs 56
4.5 Other annotation research 57
4.5.1 Qur'anic Concept Ontology 57
4.5.2 Annotation of intonation and pronunciation (*Tajweed*) 58
4.6 Qur'an as a Corpus 59

4.6.1 Sampling and Representativeness 60
4.6.2 Finite Size 62
4.6.3 Machine-readable form 62
4.6.4 Standard Reference 63
4.7 Summary 63
Chapter 5 QurAna: Corpus of the Qur'an annotated with Pronominal Anaphora 64
5.1 Introduction 64
5.2 Annotation Scheme 64
5.2.1 Related annotation schema 64
5.2.2 QurAna Annotation Schema 66
5.3 Annotation Process 67
5.4 Concept Ontology from QurAna 68
5.5 Quality Assurance 70
5.6 Applications 71
5.7 Quantitative Measures of QurAna 72
5.8 Challenges and future improvement 73
5.9 Summary 75
Chapter 6 QurSim: A corpus for evaluation of relatedness in short texts 76
6.1 Introduction 76
6.2 Book of Tafsir by Ibn Katheer 77
6.3 Compilation process 77
6.3.1 Dataset filtration and extension 78
6.4 Quality Assurance 81
6.5 Applications 82
6.6 Challenges 83
6.7 Future Improvements 85
6.8 Summary 86
Chapter 7 Text Mining Application on the Qur'an 87
7.1 Qur'anic Concordancer: QurConcord 87
7.2 QurAna: Qur'anic Pronoun Referents Application 89
7.3 Qur'anic Verse Similarity and Relatedness Application 92
7.3.1 Lexical Similarity 92
7.3.1.1 PHP Similar_Text function 92
7.3.1.2 Text::Similarity::Overlaps Module 93

7.3.1.3 Vector Similarity using TF-IDF 94
7.3.1.4 Verse segment lexical similarity 96
7.3.2 Semantic Relatedness through QurSim 98
7.3.2.1 Graph of network of verses 98
7.3.2.2 directly and indirectly related verses 99
7.3.2.3 concept clouds 101
7.3.3 Semantic relations between Qur'anic chapters 101
7.4 n-Gram Search 102
7.5 Qur'anic Word Co-Occurrence 104
7.6 Color coded POS display 105
7.7 QurCloud: Qur'anic Chapter Word Cloud 105
7.8 Concept ontology of the Qur'an 107
7.9 Summary 108
Chapter 8 Machine Learning Experiments 109
8.1 Introduction 109
8.2 Meccan and Medinan Chapters 109
8.2.1 Significance 109
8.2.2 Source of Information 110
8.2.3 Features of Meccan and Medinan Chapters 113
8.2.4 Searching resources 115
8.2.5 Counting Features 115
8.3 Running Experiments using The WEKA Tool 120
8.3.1 ARFF file 121
8.3.2 WEKA Explorer 122
8.3.2.1 Pre-process panel 123
8.3.2.2 Visualize Panel 124
8.3.2.3 Classify Panel 126
8.3.2.4 Clustering 131
8.4 Re-running experiments using QurAna dataset 135
8.5 Experiments for calculating verse distance using QurAna and QurSim 140
8.6 Summary 142
Chapter 9 Conclusion and Future Work 144
9.1 Overview 144
9.1.1 Overall findings 144
9.1.2 Chapter Summaries 144

Chapter 1 144
Chapter 2 145
Chapter 3 145
Chapter 4 146
Chapter 5 146
Chapter 6 147
Chapter 7 147
Chapter 8 147
9.2 Aims and Objectives 148
9.3 Future works 149
9.3.1 Improvements on QurAna and QurSim 149
9.3.2 Potential Machine Learning applications to Qur'anic Studies 150
9.3.2.1 Discovering patterns from Prophet's companion's exegesis 151
9.3.2.2 Discover patterns and correlations among the seven readings of the Qur'an 151
9.3.2.3 Machine Learning on Syntactic and Linguistic Patterns: 152
9.3.2.4 Machine Translation 153
9.3.2.5 Computational Stylistics 154
9.3.2.6 Subjectivity Analysis 154
Bibliography 156
List of Abbreviations 168
Appendix A Header of the download file for QurSim Dataset 169
Appendix B Header of the download file for QurSim Dataset 171

List of Tables

Table 2.1– Arabic Pronominal System 18

Table 5.1 - 20 Most frequent concepts in the Qur'an 69

Table 5.2 – Quantitative measures from QurAna corpus..... 73

Table 7.1 – comparison of different similarity measures against verse 27:65..... 98

Table 8.1 - Meccan (K) and Medinan (D) sūra index. (*) indicates a debatable case. 111

Table 8.2 – Summary of features used to classify Meccan and Medinan Chapters 120

Table 8.3- Analysis of the outcome of J48 classification of the 21 debatable suras. 130

Table 8.4 – Count of Prophet names in the Qur'an addressed as pronouns..... 135

Table 8.5 – Count of eschatological terms in the Qur'an addressed as pronouns..... 136

Table 8.6 - Count of 'marriage' related terms in the Qur'an addressed as pronouns..... 136

Table 8.7 - Count of the concepts related to 'People of the Book' in the Qur'an addressed as pronouns..... 137

Table 8.8 - Count of the concepts related to 'pillars of Islam' in the Qur'an addressed as pronouns 137

Table 8.9 – Attribute counts of 21 test cases for Meccan-Medinan classification. Prediction-1 shows classifier result before QurAna counts, and Prediction-2 is after incorporating QurAna counts. Shaded cells show the misclassification instances. Shaded columns are the attributes where QurAna was incorporated. 140

List of Qur'anic Verses

Qur'anic Verse - 23:67 ..... **5**
Qur'anic Verse - 2:23 ..... **10**
Qur'anic Verse - 17:88 ..... **10**
Qur'anic Verse - 10:38 ..... **10**
Qur'anic verse - 1:6,7 ..... **11**
Qur'anic Verse – 4:69 ..... **11**
Qur'anic Verse – 2:14 ..... **13**
Qur'anic Verse - 19:4 ..... **13**
Qur'anic Verse - 33:10 ..... **13**
Qur'anic Verse - 12:82 ..... **14**
Qur'anic Verse - 5:75 ..... **14**
Qur'anic Verse - 2:197 ..... **14**
Qur'anic Verse - 5:2 ..... **14**
Qur'anic Verse - 2:124 ..... **19**
Qur'anic Verse - 5:8 ..... **19**
Qur'anic Verse - 97:1 ..... **19**
Qur'anic Verse - 112:1 ..... **19**
Qur'anic Verse - 33:35 ..... **20**
Qur'anic Verse - 17:105 ..... **20**
Qur'anic Verse - 35:9 ..... **21**
Qur'anic Verse - 10:22 ..... **21**
Qur'anic Verse - 2:17 ..... **21**
Qur'anic Verse – 17:106 ..... **22**
Qur'anic Verse – 39:23 ..... **22**
Qur'anic Verse – 22:29 ..... **23**
Qur'anic Verse – 3:96 ..... **23**
Qur'anic Verse – 81:17 (literal translation) **23**
Qur'anic Verse - 74:33 ..... **23**
Qur'anic Verse – 91:4 ..... **23**
Qur'anic Verse - 93:2 ..... **23**
Qur'anic Verse - 2:7 ..... **24**
Qur'anic Verse - 45:23 ..... **24**

Qur'anic Verse - 2:37 24
Qur'anic Verse - 7:23 24
Qur'anic Verse - 5:1 25
Qur'anic Verse - 5:3 25
Qur'anic Verse - 6:152 25
Qur'anic Verse – 4:6 26
Qur'anic Verse – 6:8 26
Qur'anic Verse - 25:7 26
Qur'anic Verse – 2:74 26
Qur'anic Verse – 5:13 26
Qur'anic Verse – 57:16 27
Qur'anic Verse – 2:51 27
Qur'anic Verse – 20:88 27
Qur'anic Verse – 1:2 27
Qur'anic Verse – 30:18 27
Qur'anic Verse – 28:70 28
Qur'anic Verse – 84:1 28
Qur'anic Verse – 55:37 28
Qur'anic Verse – 69:16 28
Qur'anic Verse – 25:25 28
Qur'anic Verse - 15:74 28
Qur'anic Verse – 51:33 29
Qur'anic Verse – 4:154 29
Qur'anic Verse – 2:65 29
Qur'anic Verse – 16:118 29
Qur'anic Verse – 6:146 30
Qur'anic Verse – 2:219 31
Qur'anic Verse – 4:43 31
Qur'anic Verse – 5:90 31
Qur'anic Verse – 16:44 32
Qur'anic Verse – 6:36 45
Qur'anic Verse – 18:13 45
Qur'anic Verse – 10:107 46
Qur'anic Verse – 28:68 46
Qur'anic Verse – 2:2 – Translation modified from Pickthall to express the pause mark significance 47

Qur'anic Verse – 11:70 56
Qur'anic Verse – 54:19 57
Qur'anic Verse – 19:11 57
Qur'anic Verse – 16:71 57
Qur'anic Verse – 4:95 57
Qur'anic Verse - 31:27 60
Qur'anic Verse - 43:3 61
Qur'anic Verse - 26: 192-195 61
Qur'anic Verse - 39:28 61
Qur'anic Verse - 25:33 61
Qur'anic Verse - 20:123 61
Qur'anic Verse – 67 : 5 66
Qur'anic Verse - 23:67 69
Qur'anic Verse - 22:15 70
Qur'anic Verse – 65:1 74
Qur'anic Verse – 67:5 74
Qur'anic Verse – 2:26 78
Qur'anic Verse – 6:44 78
Qur'anic Verse – 78:20 79
Qur'anic Verses – 20:105-107 79
Qur'anic Verse - 78:20 80
Qur'anic Verse - 18:47 80
Qur'anic Verse - 52:10 80
Qur'anic Verse – 2:255 84
Qur'anic Verse – 11:97 84
Qur'anic Verses – 79:21-26 85
Qur'anic Verse - 2:255 97
Qur'anic Verse – 2:219 110
Qur'anic Verse – 5:90 110
Qur'anic Verse – 27:66 142
Qur'anic Verse -5:6 152

List of Figures

Figure 3.1 – Chapter clusters reported by (Thabet 2005) 35

Figure 3.2 – Chapter clusters reported by (Moisl 2009) 37

Figure 3.3 – A number of Qur'anic chapter chronology systems reported by (Sadeghi 2011) 37

Figure 4.1 – Query Interface for Haifa database (Dror et al 2004) 50

Figure 4.2 – Morphology analysis for word 11:28:16 from QAC 50

Figure 4.3 – Syntactic annotation for verse 1:1 from QADT 52

Figure 4.4 - Pronoun resolution of verse 10:58 53

Figure 4.5 – Semantic frames for the concept 'praise' in the Qur'an ... 56

Figure 4.6 - partial concept graph from QAC showing the category 'physical substance' 58

Figure 4.7 – excerpt from (Taha 2008) showing colour coded annotation of tajweed rules 59

Figure 5.1 – Example of annotation scheme from UCREL project (Garside et al. 1997) 64

Figure 5.2 – Example of annotation scheme from MUC-7 SGML schema (Hirschman and Chinchor 1997) 65

Figure 5.3 – Example of annotation scheme from GNOME project (Poesio 2004) 65

Figure 5.4 - Example of annotation scheme from AQA project (Boldrini et al. 2009) 66

Figure 5.5 - Example of annotation scheme of Arabic anaphora corpus by (Hammami et al 2009) 66

Figure 5.6 – Excerpt from verse 67:5 showing QurAna annotation schema 67

Figure 5.7 - Pronoun resolution of verse 38:29 72

Figure 6.1 – A sample QurSim XML representation 81

Figure 6.2 - Verses directly related to 7:187 83

Figure 6.3 - Concept cloud from pronoun referents of all related verses to 7:187 83

Figure 7.1 – First 10 concordance lines for the input word "كتاب" 87

Figure 7.2 – sample concordance lines for the verb 'eat' in the Qur'an (Muhammad and Atwell 2009) 89

Figure 7.3 - Pronoun resolution of verse 7:25 90

Figure 7.4 - concordance lines for the concept 'children of Adam' 91

Figure 7.5 – Lexical similarity using PHP `Similar_Text()` function 93

Figure 7.7 – lexical similarity using vector similarity (TF*IDF) 96

Figure 7.8 – Network of verses related directly or indirectly to verse 27:65 from QurSim dataset 99

Figure 7.9 – verses directly related to verse 27:65 from QurSim 100

Figure 7.10 – Concept cloud from pronouns of all verses related to verse 27:65 101

Figure 7.11 - Relatives of chapter No. 2 "Al-Baqarah", red nodes are Meccan chapters, whereas green nodes represent Medinan chapters 102

Figure 7.12 – The top 5 n-gram patterns for the word Allah with their frequencies. 103

Figure 7.13 – Top 5 results from the n-gram (من دون الله) "other than Allah" 103

Figure 7.14 – Co-occurrences of the Qur'anic word (سماء) "sky/heaven" 104

Figure 7.15 – color coded POS display of chapters 1 and 114 105

Figure 7.16 – Word cloud from chapters 113 and 114 106

Figure 7.17 – first few concepts from Qur'anic pronouns. The number refers to the count of verses under each concept. 108

Figure 8.1 - WEKA Explorer pre-processing tab. 123

Figure 8.2 – Visualizing all attributes in WEKA 124

Figure 8.3 - Visualize pane 125

Figure 8.4 – comparing two attributes 125

Figure 8.5 - WEKA classification on 93 chapters from the training set 126

Figure 8.6 - Decision tree using J48 classifier on 93 chapters 127

Figure 8.7 - Decision tree produced by 'random tree' classifier 127

Figure 8.8 - Decision tree based on ADT classifier. Note that negative values indicate Meccan and positive values indicate Medinan 128

Figure 8.9 - Check 'output predictions' for machine prediction on the 21 test chapters 128

Figure 8.10 - Outcome of the J48 Machine Learning algorithm on the 21 debatable chapters. 129

Figure 8.11 - K-Means clustering into 2 clusters 132

Figure 8.12 - Cluster of the 114 sūra into two clusters against 'kalla' attribute 133

Figure 8.13 - Clustering based on EM algorithm 134

Chapter 1
Introduction

This research contributes to the area of corpus annotation and text mining. The objective is to create a platform that enables analysis of unstructured information in search of interesting knowledge embedded within the text. The raw text is usually pre-processed and stored into intermediate representation that contains extra annotated information. The text mining application then performs a series of operations on this intermediate representation to fulfil a user's request and produce results in a visually appealing format. Text mining thus returns to users information that would otherwise be impossible to discover –or would take a long time- through manual processing. A number of underlying disciplines interact in the background to collectively cater for text mining needs, prominent among them are Natural Language Processing (NLP), Machine Learning (ML) and Information Retrieval (IR).

Although text mining applications can work on unrestricted text, practical and efficient solutions restrict their domain. For example, biomedical text mining solutions offer a wide range of useful applications restricted to the electronic literature available under the biomedical and molecular biology domain. This research restricts its domain to Islamic religious scripture where novel annotations and text mining applications are created on the original Qur'an in classical Arabic. This work can be extended to include other texts in the same domain, like texts of Hadith – i.e., sayings and traditions of Prophet Muhammad, or even religious texts in Modern Standard Arabic. Detailed discussion of the rationale behind choosing this text and domain is given in section 2.1.

1.1 Proposed Text Mining Solution

In order to enable text mining applications on the Arabic Qur'an, a number of pre-processing steps were undertaken and annotation information added to the Qur'an. The raw Arabic Qur'an was pre-processed into morphological units using the Qur'anic Arabic Corpus (QAC). Qur'anic terms were indexed and converted into a vector space model using techniques in IR. In parallel, nearly 24,000 Qur'anic personal pronouns were annotated with information on their referents. These referents are consolidated and organized into a total of over 1,000 ontological concepts. This corpus is named **QurAna** – or **Qur'an**ic **Ana**phora

Corpus (see chapter 5). Moreover, a dataset of nearly 8,000 pairs of related Qur'anic verses are compiled from books of scholarly commentary on the Qur'an. This dataset is given the name **QurSim** – or **Qur'an**ic **Sim**ilarity Corpus (see chapter 6). The vector space model, the QurAna, the QurSim, and the part-of-speech tags available in QAC all together served for a number of Qur'anic text mining applications which were rendered online for public use. Among these applications: lemma concordance, collocation, POS search of the Qur'an, verse similarity measures, concept clouds of a given verse, pronominal anaphora and Qur'anic chapter similarity (see chapter 7).

Furthermore, machine learning experiments were conducted on automatic detection of verse similarity/relatedness as well as categorization of Qur'anic chapters based on their chronology of revelation into Meccan and Medinan chapters (see chapter 8). Specialized domain attributes and linguistic features were investigated to induct learning algorithms. Results show that in order for machine learning algorithms to reach the human upper bound in certain computational tasks on Qur'anic text, it requires deep linguistic, contextual and external world knowledge. Mostly, these are tasks that require more semantic and discourse analysis, for example, verse relatedness, question answering and textual entailment. However, many useful queries on the Qur'an - discussed by domain scholars since early times- can be addressed using text mining techniques relying on the layers of annotations made available through this research.

1.2 Novel Contributions of this thesis

To the best of my knowledge, no full-fledged computational PhD programme has been conducted earlier on Qur'anic annotation and text mining. The novelty of this research lies in attempting to apply techniques from computational linguistics to a novel domain of the Qur'anic text. Many NLP techniques in Arabic or English NLP were needed to be adapted for the Qur'anic texts. This novel research has motivated other researchers towards the subject, and very often, I receive request from NLP research community requesting for collaboration and extension on this work.

The novel contribution of this thesis to computational Qur'anic studies can be summarized as follows:

- Adding a novel annotation layer on the entire Qur'an, i.e., annotation of Qur'anic personal pronouns with their antecedents. To the best of our knowledge, no pronominal annotation on the entire Qur'an exists. See chapter 5 for details.
- Compiling a dataset of semantically related verses from a scholarly source, and converting it into machine readable representation that enables computational experimentation. See chapter 6 for details.
- Compiling an ontological map of Qur'anic concepts emanating from pronoun referents. Although several topical indexes of the Qur'an exist, the ontology I developed is unique in being grounded on attachment of concepts with each pronoun, and exhaustively covering all the Qur'an. It has been found that the Qur'an exhibits very frequent use of pronouns, which rationalised this initiative to add an ontological map out of pronouns tagging. Existing ontologies and subject maps of the Qur'an can be extended to incorporate our ontology. See section 5.4 for details.
- Designing and implementing online text mining applications that enables corpus based language analysis of the Qur'an. This includes: lemma concordance, collocation, POS search of the Qur'an, verse similarity measures, concept clouds of a given verse, pronominal anaphora and Qur'anic chapter similarity. See chapter 7 for details.
- Conducting machine learning experiments utilizing Qur'anic Corpus and QurAna and QurSim datasets for a Qur'anic studies classification problem. See chapter 8 for details.

1.3 Potential Usage of the resources produced by this thesis

Further to the novel contribution of this thesis, following is a description of other potential contributions this thesis could add to a number of relevant fields, among them: Arabic NLP, computational text similarity analysis, computational anaphora resolution, stylometrics, and translation studies.

1.3.1 Arabic NLP

As the Qur'anic text does not vary much in morphology and syntax from standard modern Arabic (see section 2.1.1), the resources contributed through this thesis will benefit the Arabic NLP community. Particularly, the corpus on anaphora resolution (i.e., QurAna) will be a very useful resource for researchers in the field

of Arabic pronoun resolution. Similarly, the dataset on related verses (i.e., QurSim) is a unique contribution for computational analysis of semantic relatedness in short Arabic texts.

Moreover, the Qur'anic text is considered to be the finest form of Arabic literature, and the resources I contributed on the Qur'an would prove very helpful for researchers of classical Arabic grammar.

1.3.2 Computational Text similarity and relatedness

There has been much research in recent years on semantic relatedness at word level. The task gets more complicated when studying semantic similarity and relatedness at phrasal and sentence level. A major bottleneck towards efficient computational solution is the availability of high-quality evaluation datasets. To this end, the QurSim dataset (see chapter 6 for details) brings relief by providing a comparatively large dataset of pairs of related short texts. As the pairs represent Qur'anic verses that are tagged with unique numbers, and as multiple Qur'anic translations are available, this dataset can easily be ported to other languages and thus could contribute as an evaluation resource for researchers in semantic similarity and relatedness in these languages. In fact, an English version of this dataset was automatically compiled and communicated to Justin Washtell for experimentation using Expectation Vectors (Washtell 2011).

1.3.3 Computational Anaphora Resolution

Pronominal Anaphora constitutes a major portion of overall anaphora in a text. Despite the many useful NLP applications an anaphora resolution system has, not many evaluation corpora exist on this subject. The QurAna corpus (see chapter 5 for details) should fill a gap in this regard. As is the case with QurSim, our annotation of Qur'anic pronouns preserves verse and word numbers thus allowing alignment of pronouns to other word-by-word literal translations of the Qur'an.

1.3.4 Computational Stylistics

Stylistics focuses on the style of the text through various textual features that uniquely identifies the author's ability to evoke feelings in the audience. The Qur'an has been a subject of such studies since its revelation. The new annotation layer I created (i.e., QurAna) will ease empirical analysis of Qur'anic pronouns for detecting pronominal stylistic features of the Qur'an.

1.3.5 Translation studies

Resolving pronoun antecedents is vital for machine translations. There are already translations of the Qur'an to many languages and often multiple translations within one language. The QurAna corpus will contribute in evaluating the quality of these translations and verifying correct resolution of pronoun antecedents.

Take for an example verse 23:67 as follows[1]:

مُسْتَكْبِرِينَ بِهِ سَامِرًا تَهْجُرُونَ
(Being) arrogant / about it, / conversing by night, / speaking evil." /
In scorn thereof. Nightly did ye rave together.

Qur'anic Verse - 23:67

The pronoun 'it' in this verse has no explicit antecedent mentioned and as such could refer to multiple entities like, Prophet Muhammad, the Qur'an or 'the house of Allah'. I have found out through empirical experiments that 54% of pronouns in the Qur'an have no antecedence. More elaboration will be given in chapter 5. For this verse and after consulting books of Tafsir, the majority opinion favours that this pronoun refers to 'the house of Allah', and as such I have tagged this pronoun accordingly in QurAna. However, when I compared the six English translations for this verse available in (Quran.com) website as listed in the box below, I notice that 2 and 4 marked the reference of the pronoun in brackets as "the Qur'an", whereas 1,3 and 6 did not included any comments on the referent. Shakir was the most non-literal and replaced the pronoun in the translation with referent. Remarkably, none of these six prominent translators suggested the referent to be "the house of Allah" which is the most appropriate one according to Ibn Katheer (Ibn-Katheer 1372).

1. Sahih International
In arrogance regarding it, conversing by night, speaking evil.
2. Muhsin Khan

[1] Unless otherwise mentioned, all English translations were taken from online rendering of (Pickthall 1973) available at http://Quran.com

In pride (they Quraish pagans and polytheists of Makkah used to feel proud that they are the dwellers of Makkah sanctuary Haram), talking evil about it (the Qur'an) by night.

3. Pickthall

In scorn thereof. Nightly did ye rave together.

4. Yusuf Ali

"In arrogance: talking nonsense about the (Qur'an), like one telling fables by night."

5. Shakir

In arrogance; talking nonsense about the Qur'an, and left him like one telling fables by night.

6. Dr. Ghali

Waxing proud against it, forsaking it for entertainment."

This example shows the value-added our corpus provides to the evaluation of existing translations and the active role it can play in building new translations in English or other languages. Moreover, this resource would be useful for authors who would attempt a novel translation of the Qur'an to a language that has no Qur'an translation done yet.

1.4 Thesis Outline

The reminder of this thesis is organized as follows. Chapter 2 is dedicated to a background introduction of the main subjects of this thesis, starting with introducing the text under investigation: the Qur'an. After a general overview of the text, a detailed discussion is given to rationalise usage of this text for text mining and NLP research. In addition, this chapter provides some linguistic background on Arabic pronominal anaphora and text relatedness, particularly focusing on classical Arabic of the Qur'an. This chapter also highlights the importance of Qur'anic translations and Qur'anic exegesis – or books of Tafsir.

Chapter 3 reviews literature on major computational works done under corpus and computational linguistics discipline related to this thesis, particularly areas under: computational Qur'anic studies, anaphora resolution and computational analysis of similarity and relatedness in short texts.

Chapter 4 reviews existing linguistic annotations of the Qur'an. Starting from the raw Qur'anic text, I reviewed available annotations done on the Qur'anic text. A detailed review of two existing corpora on the Qur'an is given: the Haifa corpus and the Qur'anic Arabic Corpus (QAC). In addition, this chapter highlights other partial attempts to annotate the Qur'an with other linguistic and semantic analysis like dependency parse trees and semantic frames. This chapter also investigates corpus features of the Qur'an.

Chapters 5 and 6 give a detailed description of the two major language resources contributed through this thesis, namely QurAna and QurSim respectively. Each chapter describes the annotation scheme, the annotation process, evaluation, quality assurance, applications and major challenges and future improvements.

Chapter 7 describes a number of text mining applications created online using the language resources discussed earlier, i.e., QAC, QurAna, and QurSim. The usefulness of each application to the domain users is emphasised by showing how these applications alleviates the pain of daunting manual tasks carried by Qur'anic researchers today.

Chapter 8 details machine learning experiments carried out on the Qur'anic text, enriching features from developed corpora. This includes experiments on categorizing Qur'anic chapters chronologically into Meccan and Medinan based on certain linguistic features. Other experiments explored computational analysis of text similarity between two verses based on vector space models.

Chapter 9 concludes this thesis summarizing the main contribution of this thesis to the user community and discussing potential fronts where this research could be extended in future.

1.5 Summary

This research focuses on building language resources that would eventually enter into beneficial text mining applications. The main contribution of this thesis has been the development of three novel resources, namely, a corpus of Qur'anic pronoun tagging, a domain concept ontology of the Qur'an emanating from

pronoun referents and a dataset of related verses in the Qur’an compiled from scholarly sources. These resources were used in a number of text mining applications and machine learning experiments.

This chapter explored a number of possible ways this research could add value to the NLP community, particularly: Arabic NLP, computational text similarity analysis, computational anaphora resolution, stylometrics, and translation studies.

Chapter 2
Background on the language of the Qur'an

2.1 What is the Qur'an?

The Qur'an is a scripture which is according to Muslims the verbatim words of Allah containing over 77,000 words revealed through Archangel Gabriel to Prophet Muhammad over 23 years beginning in 610 CE. It is divided into 114 chapters of varying sizes, where each chapter is divided into verses, adding up to a total of 6,243 verses. The Qur'anic chapters (known as Suras) are classified into Meccan or Medinan depending upon the time of their revelation in respect to the Hijra or the Migration of Prophet from Makkah to Medina. The two categories of chapters vary in thematic as well as stylistic approaches. More detailed analysis of these two categories is made in chapter 8.2.

2.1.1 Classical Arabic (CA) vs. Modern Standard Arabic (MSA)

The form of the Arabic language used in the Qur'an is often called the Classical Arabic (CA), which is the form of Arabic language used in literary texts authored by early Arabic scholars from the 6th through 10th century. In contrast to most modern languages, the published body of text in the Classical Arabic is very rich and both in size and language usage. The corpora of Modern Standard Arabic (MSA) - the form used in contemporary scholarly works as well as in the media- on the other hand is growing in size over time and new technical terms are being introduced, but the richness of Arabic vocabulary and stylistic usage is superior in the Classical Arabic register. For example, Maha Al-Rabiah has compiled a corpus of Classical Arabic of 50 million words from the period between seventh and eleventh Gregorian century (Al-Rabiah 2012) comprising of -in addition to the Qur'an- reference materials from a wide range of topics including: religion, linguistics, literature, science, sociology and biography. MSA does not differ from Classical Arabic in morphology or syntax, but richness of stylistic and lexis usage is apparent in Classical works. Thus, both CA and MSA could be thought of as two registers of the same language. Given this, I believe conducting computational and linguistic research work on CA would benefit also the MSA. However, most recent work on Arabic corpus annotation has concentrated on MSA, and the computational corpus linguistic community has largely ignored the study of the language resources available in Classical Arabic. This has been one of the motivations behind this thesis (Muhammad and Atwell 2012a).

2.1.2 Inimitability of the Qur'an

The Qur'an was revealed to Prophet Muhammad in verbal form, and so it continued to the death of the Prophet when Caliph Abu Bakr commissioned a team of scholars to record the Qur'an in a written format collecting from scattered sources, thus guarding it against getting lost. Two decades later (651 CE), when Islam spread to foreign lands, Caliph Uthman produced a standard codex for the written form out of this master copy, and distributed copies to all distant lands.

Being words of God, Qur'an is considered inimitable by Muslims. Muslim Scholars consider the linguistic style of the Qur'an unique and cannot be imitated by human based on a number of Qur'anic verses on this matter as follows. In fact, in multiple verses, the Qur'an has placed challenge to mankind to produce a book like this Qur'an (e.g., verses 2:23, 17:88) or even a chapter like it (e.g., verse 10:38).

وَإِن كُنتُمْ فِي رَيْبٍ مِّمَّا نَزَّلْنَا عَلَىٰ عَبْدِنَا فَأْتُوا بِسُورَةٍ مِّن مِّثْلِهِ وَادْعُوا شُهَدَاءَكُم مِّن دُونِ اللَّهِ إِن كُنتُمْ صَادِقِينَ

And if ye are in doubt concerning that which We reveal unto Our slave (Muhammad), then produce a surah of the like thereof, and call your witness beside Allah if ye are truthful.

Qur'anic Verse - 2:23

قُل لَّئِنِ اجْتَمَعَتِ الْإِنسُ وَالْجِنُّ عَلَىٰ أَن يَأْتُوا بِمِثْلِ هَٰذَا الْقُرْآنِ لَا يَأْتُونَ بِمِثْلِهِ وَلَوْ كَانَ بَعْضُهُمْ لِبَعْضٍ ظَهِيرًا

Say: Verily, though mankind and the jinn should assemble to produce the like of this Qur'an, they could not produce the like thereof though they were helpers one of another.

Qur'anic Verse - 17:88

أَمْ يَقُولُونَ افْتَرَاهُ ۖ قُلْ فَأْتُوا بِسُورَةٍ مِّثْلِهِ وَادْعُوا مَنِ اسْتَطَعْتُم مِّن دُونِ اللَّهِ إِن كُنتُمْ صَادِقِينَ

Or say they: He hath invented it? Say: Then bring a surah like unto it, and call (for help) on all ye can besides Allah, if ye are truthful.

Qur'anic Verse - 10:38

2.1.3 Some Linguistic features

The Qur'an being part of the Classical Arabic register, exhibits rich features of language usage. Following is a brief discussion of some linguistic and rhetorical features of the Qur'an. These features are not particularly unique to the Qur'an

as a text, however, their extent of usage is more prominent and significant in the Qur'an. Most information in this section was adapted from our published paper (Muhammad and Atwell 2009)

2.1.3.1 Scattered information on the same topic

The Qur'an often talks about a topic scattered within many different verses in different chapters. Consider the following two sets of verses. In 1:6,7 there is a reference to a 'straight/right path' and a reference to a category of people whom God has favoured without highlighting who might be in this category. Verse 4:69 elaborates on this query and gives examples of four categories whom Allah has favoured .

اهْدِنَا الصِّرَاطَ الْمُسْتَقِيمَ صِرَاطَ الَّذِينَ أَنْعَمْتَ عَلَيْهِمْ
Show us the straight path, The path of those whom Thou hast favoured

Qur'anic verse - 1:6,7

وَمَن يُطِعِ اللَّهَ وَالرَّسُولَ فَأُولَٰئِكَ مَعَ الَّذِينَ أَنْعَمَ اللَّهُ عَلَيْهِم مِّنَ النَّبِيِّينَ وَالصِّدِّيقِينَ وَالشُّهَدَاءِ وَالصَّالِحِينَ ۚ
Whoso obeyeth Allah and the messenger, they are with those unto whom Allah has shown favour, of the prophets and the saints and the martyrs and the righteous

Qur'anic Verse – 4:69

The Qur'an also repeats a story, for example, stories of prophets. However each instance of repetition adds some details not available in other instances. For example, the Qur'an tells various aspects of the story of Moses in 132 places distributed among 20 chapters. It is one of the novel contributions of this thesis to study verse relatedness and compile a dataset for future computational analysis. See chapter 6 for details.

2.1.3.2 Literal vs. technical sense of a word

The Qur'an borrows an existing Arabic word and specializes it to indicate a technical term. Consider for example the word جَنَّة */jannah* meaning literally 'a garden', but whenever this word is used in the Qur'an, it refers to 'the paradise' where the believers will abode as reward after the Day of Judgment. However, there are some instances where this word is used in its original literal meaning to

refer to certain gardens in this world. In the following two verses, 3:133 uses the more frequent technical sense and 34:15 uses the less frequent literal meaning.

وَسَارِعُوا إِلَىٰ مَغْفِرَةٍ مِّن رَّبِّكُمْ وَجَنَّةٍ عَرْضُهَا السَّمَاوَاتُ وَالْأَرْضُ أُعِدَّتْ لِلْمُتَّقِينَ
And vie one with another for forgiveness from your Lord, and for a paradise as wide as are the heavens and the earth, prepared for those who ward off (evil);

Qur'anic Verse - 3:133

لَقَدْ كَانَ لِسَبَإٍ فِي مَسْكَنِهِمْ آيَةٌ ۖ جَنَّتَانِ عَن يَمِينٍ وَشِمَالٍ
There was indeed a sign for Sheba in their dwelling-place: Two gardens on the right hand and the left.

Qur'anic Verse - 34:15

2.1.3.3 Verbs preposition binding

The Qur'an exhibits many examples where a certain verb is associated with a preposition which is unusual with this verb, but common with a different verb. Consider verse 2:14 with two translations [a] and [b] below, the Arabic verbs خلا/*khala* means "be alone", which is usually followed by the preposition 'with' like 'John was alone with Mary'. However, in this verse the Qur'an choose to use the preposition 'to' with 'be alone' which sounds unusual to say, 'John was alone to Mary'. However, this is a valid classical Arabic style when a verb borrows a preposition that binds with another verb and uses it to indicate at the same time the meaning of both verbs. The Arabic verb ذهب/*dhahaba* (go) fits well with the preposition 'to' as in: 'John went to Mary'. So, in this verse, the Qur'an by using a verb (be alone) with a preposition (to) from another verb 'go' conveyed the meaning of both "being alone" and "going to" at the same time. This unique characteristic made both translations in [a] and [b] partially true, highlighting either the sense of the original verb 'be alone with' as in [a] or the implicit verb with explicit preposition 'go to' as in [b]. See (Ibn-Katheer 1372) on his commentary on this verse.

وَإِذَا لَقُوا الَّذِينَ آمَنُوا قَالُوا آمَنَّا وَإِذَا خَلَوْا إِلَىٰ شَيَاطِينِهِمْ قَالُوا إِنَّا مَعَكُمْ إِنَّمَا نَحْنُ مُسْتَهْزِئُونَ
[a] When they meet those who believe, they say: "We believe;" but when they are alone with their evil ones, they say: "We are really with you: We (were) only jesting." [2:14 Yusuf Ali Translation]

Yusuf Ali Translation

[b] And when they fall in with those who believe, they say: We believe; but when they go apart to their devils they declare: Lo! we are with you; verily we did but mock.

Pickthall Translation

Qur'anic Verse – 2:14

2.1.3.4 Metaphor and Figurative use

The Qur'an uses a lot of metaphors and figurative language. Below are two examples. In 19:4 Pickthall translation used the verb 'shine' but the Arabic verb اشتعل/*ishtala* means 'to flare' and compares spread of gray hair with a 'fire burning a bush'. In 33:10, the Muslim army was so frightened that it felt as if their "hearts reached to their throats".

قَالَ رَبِّ إِنِّي وَهَنَ الْعَظْمُ مِنِّي وَاشْتَعَلَ الرَّأْسُ شَيْبًا

My Lord! Lo! the bones of me wax feeble and my head is shining with grey hair.

Qur'anic Verse - 19:4

إِذْ جَاءُوكُم مِّن فَوْقِكُمْ وَمِنْ أَسْفَلَ مِنكُمْ وَإِذْ زَاغَتِ الْأَبْصَارُ وَبَلَغَتِ الْقُلُوبُ الْحَنَاجِرَ

When they came upon you from above you and from below you, and when eyes grew wild and hearts reached to the throats

Qur'anic Verse - 33:10

2.1.3.5 Metonymy

In many verses the Qur'an uses metonymy, which is a rhetorical tool whereby a thing is not called by its own name but rather by something associated with that name. For example, in 12:82 the Arabic verse literally means 'ask the town' which was intended to be (and was translated so) 'ask the people who live in the town'. In 5:75 'eating food' metonymically symbolizes a human nature that contrast with god's who does not need provision, see for example (Ibn-Katheer 1372) commenting on this verse.

وَاسْأَلِ الْقَرْيَةَ الَّتِي كُنَّا فِيهَا وَالْعِيرَ الَّتِي أَقْبَلْنَا فِيهَا ۖ

Ask the township where we were, and the caravan with which we travelled hither.

Qur’anic Verse - 12:82

مَّا الْمَسِيحُ ابْنُ مَرْيَمَ إِلَّا رَسُولٌ قَدْ خَلَتْ مِن قَبْلِهِ الرُّسُلُ وَأُمُّهُ صِدِّيقَةٌ ۖ كَانَا يَأْكُلَانِ الطَّعَامَ

The Messiah, son of Mary, was no other than a messenger, messengers (the like of whom) had passed away before him. And his mother was a saintly woman. And they both used to eat food

Qur’anic Verse - 5:75

2.1.3.6 Imperative vs. non-Imperative senses

Arabic verb-forms or tenses are classified into past, present and imperative. Future tense in Arabic is usually denoted by prefixing a present form with a future particle like Seen (سـ) or Sawfa (سوف). The imperative sense can be understood from the type of the verb used. Although this general rule applies in the Qur’an, however there are many instances where an imperative is understood although no imperative form of verb is used. For example, verse 2:197 makes it imperative for pilgrims to abstain from lewdness during pilgrimage, however no imperative verb form is used. The opposite is also true: there are instances where an imperative verb is used, but that does not employ imperative instruction, for example, verse 5:2 uses imperative form “go hunting”, however it indicates the permission to hunt after leaving sacred territory where hunting was forbidden. Note how the translator (Pickthall) explicitly indicated this non-imperative meaning within brackets (i.e., ‘if ye will’).

فَمَن فَرَضَ فِيهِنَّ الْحَجَّ فَلَا رَفَثَ وَلَا فُسُوقَ وَلَا جِدَالَ فِي الْحَجِّ

[14] and whoever is minded to perform the pilgrimage therein there is no lewdness nor abuse nor angry conversation on the pilgrimage.

Qur’anic Verse - 2:197

وَإِذَا حَلَلْتُمْ فَاصْطَادُوا ۚ

But when ye have left the sacred territory, then go hunting (if ye will).

Qur’anic Verse - 5:2

2.1.4 Why computational analysis of the Qur’an?

We considered the Qur’anic text to be a potential subject for computational analysis. In what follows few rationale behind this choice is given. (Atwell et al. 2011)

- Distribution of a particular concept or subject over many scattered verses within different chapters is very evident in the Qur'an. Often a concept summarized in one verse is elaborated in another verse. Historical events, stories of prophets, emphasis on a command, attributes and qualities of Gods, description of paradise and hell fire, are some of the common subjects that are often repeated in the Qur'an. However, each repetition adds new meanings absent in other instances, and the overall subject could be fully understood when all instances are taken into consideration. This property of the Qur'an made it an attractive text for the purpose of analysing semantic relatedness between short texts, which in our case are between individual verses, or a group of verses of the Qur'an. See chapter 6 for details.
- The original Arabic Qur'an is characterized by very frequent use of anaphors. The majority of anaphoric devices in the Qur'an appear around pronominal anaphora. Hence, the ability to resolve pronoun antecedence is vital to understanding the Qur'an. See chapter 5 for details.
- The Qur'anic scripture is a widely used and cited document that guides the lives of over 1.6 billion Muslim adherents today (according to a Pew Forum report[1]). Increasingly non-Arabic speaking Muslims –and many non-Muslims- learn classical Arabic with the objective of understanding the Qur'an. For Arabic speakers, the Qur'an is considered to be the finest piece of literature in the Arabic Language. Producing language evaluation resources for computational analysis of such an important text should be well justified. Moreover, there have been a number of language resources around other spiritual scriptures like the Bible (Resnik et al., 1999).
- The majority of works done by NLP community are concentrated on Modern Standard Arabic. Classical Arabic –which is the language of the Qur'an- consists a huge register of the Arabic literature (See sec. 2.1.1). Computational studies of the Qur'an is a first step toward investigating this register.
- Being a central text in Arabic, over the past 14 centuries a huge body of scholarly commentary volumes has been compiled elaborating on linguistic, stylistic, semantic and other aspects of the Qur'an. This makes the task of compiling evaluation datasets and annotated corpora on the

[1] "The Future of the Global Muslim Population", Pew Analysis, Janurary 27, 2011. (http://www.pewforum.org/The-Future-of-the-Global-Muslim-Population.aspx)

Qur'an simpler; as it is usually possible to find scholarly comments on any difficult annotation question. In the QurSim related-verse dataset, I relied on the Qur'anic commentary work of Ibn Katheer -a well-known Qur'anic scholar who died in 1373 CE. See chapter 6.

- The Qur'an is widely available translated into almost all live languages of the world, and in many cases there are multiple translations within one language. Among these translations a number of them are also available in machine readable electronic format in the web . All translations maintain chapter and verse numbers, allowing alignment between these translations at verse level. Moreover, to be faithful to the words of God, all translations are made very carefully. Given this fact, any language evaluation resource in the source language of the Qur'an could be used to evaluate computational tasks on the target language translations as well.
- The size of the Qur'an is manageable for manual or semi-automatic annotation tasks. Given that the Arabic language still lacks many NLP resources available for a language like English (e.g., taggers, parsers, Wordnet, frameNet, etc.), developing a manually annotated language resource on a small scale like the Qur'an text could be a good starting point. A case in point is the Qur'anic Arabic Corpus (QAC) project (Dukes et al 2010), where every word of the Qur'an is tagged with morphological, part-of-speech and syntactic information, and is publicly available for research purposes. Another available resource is QurAna, pronominal anaphora of the entire Qur'an tagged with pronoun antecedent references (Muhammad & Atwell, 2012) . Our QurSim dataset, along with these other available resources on the Qur'an, will enable interesting computational linguistic applications on the Qur'an which in turn will eventually motivate wider applications and resource development for Classical and Modern Standard Arabic.
- In Western media and WWW, 'Islam' is often associated with 'terrorism'[1]. An impartial online Qur'an text mining tool would enable Western schools, universities and general public to understand Islam and its most central source, i.e., the Qur'an.

[1] See for example, an article on the Telegraph published on 07 – June- 2010 titled "Islam associated with terrorism by public, poll shows".

- Going beyond simple ‘factoid’ question answering system, one needs deep annotation of multiple layers, and integration with external knowledge sources.
- It is important that answers returned by text mining systems on the Qur’an are logically consistent and correct. Although many NLP experiments, especially those requiring textual entailment and inferences may tolerate acceptable error margin; for the Qur’an users, any false inference may be unacceptable.

2.2 Pronominal Anaphora

Anaphora is generally meant to be any expression that receives interpretation by something mentioned before in the discourse known as the antecedent. When this anaphoric expression is restricted to be personal pronouns – which is the focus of this research – then we have Pronominal Anaphora.

Pronouns can fall into four classes: personal pronouns (e.g,: I, we, us, our, you, your, he, she, it, they, her, him, it, them), possessives (e.g., his, her, their), reflexives (e.g., himself, herself) and demonstratives (e.g., this, that, those, these). In terms of the usage, personal pronouns represent the most important class and is the focus of this research.

A detailed discussion of linguistic aspects of anaphora and pronouns is beyond the scope of this thesis. In what follows is a brief introduction to the pronominal system in Arabic, followed by a discussion of pronominal anaphora in the Qur’an.

2.2.1 Pronominal System in Arabic

Pronouns in Arabic take various forms depending on multiple morpoho-syntactic features like, person, number, gender, grammatical case (i.e., nominative, accusative or genitive), and whether the pronoun form is a separate word or suffixed to another part of speech. Table 2.1 gives a summary of all these classifications.

Note that unlike English, Arabic has dedicated dual pronouns for 2nd and 3rd person. Also, 2nd and 3rd person has different gender form to address plurals. Arabic pronouns do not dedicate a separate form for 3rd person neuter.

person	number	gender	Separate		Enclitic		
			Acc.	Nom.	Gen.	Acc.	Nom.
1st	**Sin.**	Mas.	Iyyaya إياي	Ana أنا	-i ـي	-i/-ni /ـي	-tu ـتُ

		Fem.				ـني	
	dual	Mas.	Iyyana إيانا	Nahnu نحن	-na ـنا		
		Fem.					
	Pl.	Mas.					
		Fem.					
2nd	**Sin.**	Mas.	Iyyaka إياكَ	Anta أنتَ	-ka ـكَ		-ta ـتَ
		Fem.	Iyyaki إياكِ	Anti أنتِ	-ki ـكِ		-ti/-i/ ـتِ/ـي
	dual	Mas.	Iyyakuma إياكما	Antuma أنتما	-kuma ـكما		-tuma/-a ـتما/ـا
		Fem.					
	Pl.	Mas.	Iyyakum إياكم	Antum أنتم	-kum ـكم		-tum/-mu ـتم/ـوا
		Fem.	Iyyakunna إياكن	Antenna أنتن	-kunna ـكن		-tunna/-na ـتن/ـن
3rd	**Sin.**	Mas.	Iyyahu إياه	Huwa هو		-hu ـه	
		Fem.	Iyyahuma إياها	Hiya هي		-ha ـها	
	dual	Mas.	Iyyahuma إياهما	Huma هما	-huma ـهما		-a/-an ـا/ـان
		Fem.					-ta/-an ـتا/ـتان
	Pl.	Mas.	Iyyahum إياهم	Hum هم	-hum ـهم		-u/-un ـو/ـون
		Fem.	Iyyahunna إياهن	Hunna هن	-hunna ـهن		-na ـن

Table 2.1– Arabic Pronominal System

A detailed treatment of Arabic pronoun system can be retrieved from a reference book like (Wright 1967).

2.2.2 Anaphora in the Qur'an

This subsection is adapted from our published paper on QurAna, (Muhammad and Atwell 2012a).

Pronouns in the Qur'an can refer to an entity like a person (as in verse 2:124), or a verbal phrase (as in 5:8).

وَإِذِ ابْتَلَىٰ **إِبْرَاهِيمَ** رَبُّهُ
And when / tried / **Ibrahim** / <u>his</u> Lord

And, when **Abraham** was tried by <u>his</u> Lord

Qur'anic Verse - 2:124

اعْدِلُوا هُوَ أَقْرَبُ لِلتَّقْوَىٰ
Be just / <u>it</u> / (is) nearer / to [the] piety
Deal justly, <u>that</u> is nearer to your duty

Qur'anic Verse - 5:8

The distance between the pronoun and its antecedents can vary greatly in the Qur'an. For instance, in chapter 2 of the Qur'an, a series of verses address the "Children of Israel" using the 2nd person plural "you". Sometimes the distance between the pronoun and the antecedent reaches to a maximum of 1062 word segments, as is the case between the pronoun "you" in verse 2:80 and its referent in verse 2:47.

Often we notice that the Qur'an deliberately omits a noun and replaces it with a pronoun when this omitted noun is obvious for a human reader. For example in verse 97:1, the pronoun "it" is understood to refer to the "Qur'an".

إِنَّا أَنزَلْنَاهُ فِي لَيْلَةِ الْقَدْرِ
Indeed, We / revealed <u>it</u> / in / (the) Night / (of) Power. /
Lo! We revealed <u>it</u> on the Night of Predestination.

Qur'anic Verse - 97:1

As expected, in most cases the antecedent –when available- comes before the pronoun, however in some cases the referent comes after the pronoun. Such cases are called "cataphor". For example, verse 112:1 uses "He" to refer to the cataphor "Allah".

قُلْ هُوَ **اللَّهُ** أَحَدٌ
Say, "<u>He</u> is **Allah** , [who is] One,

Qur'anic Verse - 112:1

The Qur'an makes very efficient use of pronouns as a handy tool to avoid repetition of concepts and entities mentioned previously. For instance, verse 33:35 talks about 10 category of people from both sexes, and towards the end

the pronoun “them” refers back to these categories in a way that would lead to cumbersome repetition otherwise.

إِنَّ الْمُسْلِمِينَ وَالْمُسْلِمَاتِ وَالْمُؤْمِنِينَ وَالْمُؤْمِنَاتِ وَالْقَانِتِينَ وَالْقَانِتَاتِ وَالصَّادِقِينَ وَالصَّادِقَاتِ وَالصَّابِرِينَ وَالصَّابِرَاتِ وَالْخَاشِعِينَ وَالْخَاشِعَاتِ وَالْمُتَصَدِّقِينَ وَالْمُتَصَدِّقَاتِ وَالصَّائِمِينَ وَالصَّائِمَاتِ وَالْحَافِظِينَ فُرُوجَهُمْ وَالْحَافِظَاتِ وَالذَّاكِرِينَ اللَّهَ كَثِيرًا وَالذَّاكِرَاتِ أَعَدَّ اللَّهُ لَهُم مَّغْفِرَةً وَأَجْرًا عَظِيمًا

Indeed, / the Muslim men / and the Muslim women, / and the believing men / and the believing women, / and the obedient men / and the obedient women, / and the truthful men / and the truthful women, / and the patient men / and the patient women, / and the humble men / and the humble women, / and the men who give charity / and the women who give charity / and the men who fast / and the women who fast, / and the men who guard / their chastity / and the women who guard (it), / and the men who remember / Allah / much / and the women who remember / Allah / Allah has prepared / for <u>them</u> / forgiveness / and a reward / great. /

Qur’anic Verse - 33:35

However, in certain places, we notice that –for the purpose of emphasis- Qur’an explicitly repeats a concept where pronoun could have been used, as is the case for example in verse 17:105, where the second “truth” could have been shorten with a pronoun. Note that although English translation –for the sake of clarify- included the second ‘it’, the Arabic verse does not contain an explicit second pronoun, and thus would not cause confusion of two consecutive pronouns it the second ‘truth’ were indeed replaced with a pronoun.

وَبِالْحَقِّ أَنزَلْنَاهُ **وَبِالْحَقِّ** نَزَلَ

And **with the truth** / We sent it down, / and **with the truth** / it descended.

With truth have We sent it down, and with truth hath it descended.

Qur’anic Verse - 17:105

2.2.2.1 Pronoun antecedent agreement in the Qur’an

By default, pronouns in Qur’an – like most of other languages- grammatically agree with their antecedents in terms of person, number and gender. However, in certain cases this default rule is violated for semantic or stylistics reasons. Following is a discussion of some of these cases.

One very common form of pronoun and antecedent disagreement in the Qur’an is the majestic plural, where Allah refers to himself in plural “We” form. For

example, the verse 17:105 above "We sent it down", or the verse 97:1 "We revealed it down during the Night of Decree", all referring to Allah.

It is a Qur'anic style to use a masculine pronoun to refer back to both. For example, refer back to verse 33:35 above, the Arabic pronoun used at the end is لهم "*Lahum*" (them) which is used for a group of men, although the 10 antecedents explicitly mentions both sexes.

Another common style of Qur'an is to deliberately make disagreement between a pronoun and its antecedent in terms of person or number. This "grammatical shift" is used for the purpose of drawing the attention of the listener. Take for example verse 35:9 below, where the verse starts talking about Allah in a third person form, and then suddenly shifts into the first person plural (i.e., majestic plural).

وَ**اللَّهُ** الَّذِي أَرْسَلَ الرِّيَاحَ فَتُثِيرُ سَحَابًا فَسُقْنَاهُ إِلَىٰ بَلَدٍ مَّيِّتٍ فَأَحْيَيْنَا بِهِ الْأَرْضَ بَعْدَ مَوْتِهَا ۚ كَذَٰلِكَ النُّشُورُ
And **Allah** it is Who sendeth the winds and they raise a cloud; then We lead it unto a dead land and revive therewith the earth after its death. Such is the Resurrection.

Qur'anic Verse - 35:9

Verse 10:22 is another example, where the verse starts addressing in 2nd person plural "you" and then shifts to using 3rd person plural "them".

حَتَّىٰ إِذَا كُنتُمْ فِي الْفُلْكِ وَجَرَيْنَ بِهِم بِرِيحٍ طَيِّبَةٍ
when ye are in the ships and they sail with them with a fair breeze

Qur'anic Verse - 10:22

Verse 2:17 below makes a shift in number from singular third person "him" to plural third person "their".

مَثَلُهُمْ كَمَثَلِ الَّذِي اسْتَوْقَدَ نَارًا فَلَمَّا أَضَاءَتْ مَا حَوْلَهُ ذَهَبَ اللَّهُ بِنُورِهِمْ
Their likeness is as the likeness of one who kindleth fire, and when it sheddeth its light around him Allah taketh away their light

Qur'anic Verse - 2:17

2.3 Text Similarity and relatedness in the Qur'an

The Qur'an testified itself that information is scattered this book and that its verses are semantically paired. See supportive verses below.

وَقُرْآنًا فَرَقْنَاهُ لِتَقْرَأَهُ عَلَى النَّاسِ عَلَىٰ مُكْثٍ وَنَزَّلْنَاهُ تَنزِيلًا

And the Qur'an / We have divided, / that you might recite it / to / the people / at / intervals. / And We have revealed it / (in) stages. /

And (it is) a Qur'an that We have divided, that thou mayst recite it unto mankind at intervals, and We have revealed it by (successive) revelation.

Qur'anic Verse – 17:106

اللَّهُ نَزَّلَ أَحْسَنَ الْحَدِيثِ كِتَابًا مُّتَشَابِهًا

Allah / has revealed / (the) best / (of) [the] statement - / a Book / (its parts) resembling each other / oft-repeated. /

Allah hath (now) revealed the fairest of statements, a Scripture consistent, (wherein promises of reward are) paired (with threats of punishment).

Qur'anic Verse – 39:23

What follows is a discussion of some examples of related Qur'anic verses where some ambiguity in one verse is resolved by investigating other related verses. The dataset of related verses that I compiled –i.e., QurSim (see chapter 6)- contains similar kind of relatedness. Following examples are adapted from (Shanqity 1973).

2.3.1 Ambiguity in resolving the sense of a noun

Sometimes we encounter an Arabic noun in the Qur'an which has multiple senses. However, with careful investigation, semantically linking the verse that contains such ambiguous noun with other related verses often resolves this confusion. Consider verse 22:29 below. The underlined Qur'anic word العتيق/ *al-ateeq* literally could have three separate meanings: a) ancient, b) someone/something emancipated from tyranny, c) and a generous person/object. Which of these three senses is applicable here? The answer to this query could be found in another related verse – 3:96 which comes to vote for the first sense which is the 'ancient' sense denoted by this house being 'the first Sanctuary'.

ثُمَّ لْيَقْضُوا تَفَثَهُمْ وَلْيُوفُوا نُذُورَهُمْ وَلْيَطَّوَّفُوا بِالْبَيْتِ الْعَتِيقِ

Then let them make an end of their unkemptness and pay their vows and go around the ancient House.

Qur'anic Verse – 22:29

إِنَّ أَوَّلَ بَيْتٍ وُضِعَ لِلنَّاسِ لَلَّذِي بِبَكَّةَ مُبَارَكًا وَهُدًى لِّلْعَالَمِينَ

Lo! the first Sanctuary appointed for mankind was that at Becca, a blessed place, a guidance to the peoples;

Qur'anic Verse – 3:96

2.3.2 Ambiguity in resolving the sense of a verb

وَاللَّيْلِ إِذَا عَسْعَسَ

And by the night as it arrives/departs.

Qur'anic Verse – 81:17 (literal translation)

Similar to nouns, some Qur'anic verbs can have multiple senses, which then can be resolved by looking into other related verses. For example, in verse 81:17 above, the underlined verb عسعس / *asasa* could literally mean either 'arrive' or 'depart', but which one is more probable?

Checking the Qur'an we find the following three related verses 74:33, 91:4 and 93:2. The first verse 74:33 favors the 'departure' sense, whereas the other two verses 91:4 and 93:2 convey the 'arrival' sense when the night comes to cover the day. Taking the majority vote, (Ibn Katheer 1372) favoured the 'arrival' sense.

وَاللَّيْلِ إِذْ أَدْبَرَ

And the night when it withdraweth.

Qur'anic Verse - 74:33

وَاللَّيْلِ إِذَا يَغْشَاهَا

And the night when it enshroudeth him,

Qur'anic Verse – 91:4

وَاللَّيْلِ إِذَا سَجَىٰ

And by the night when it is stillest,

Qur'anic Verse - 93:2

2.3.3 Ambiguity in resolving the sense of a particle

خَتَمَ اللَّهُ عَلَىٰ قُلُوبِهِمْ وَعَلَىٰ سَمْعِهِمْ ۖ وَعَلَىٰ أَبْصَارِهِمْ غِشَاوَةٌ
Allah hath sealed their hearing and their hearts, and on their eyes there is a covering. Theirs will be an awful doom.

Qur'anic Verse - 2:7

The conjunction particle 'and' in the Arabic text of verse 2:7 above is used to apply "anding" over three senses: "heart feeling, hearing and vision" which makes it ambiguous if sealing or veiling is applied on these senses. However, the related verse 45:23 below makes it clear that the seal is applicable over the "heart feeling and hearing", while 'covering' is applied only over "their vision".

أَفَرَأَيْتَ مَنِ اتَّخَذَ إِلَٰهَهُ هَوَاهُ وَأَضَلَّهُ اللَّهُ عَلَىٰ عِلْمٍ وَخَتَمَ عَلَىٰ سَمْعِهِ وَقَلْبِهِ وَجَعَلَ عَلَىٰ بَصَرِهِ غِشَاوَةً فَمَن يَهْدِيهِ مِن بَعْدِ اللَّهِ ۚ أَفَلَا تَذَكَّرُونَ
Hast thou seen him who maketh his desire his god, and Allah sendeth him astray purposely, and sealeth up his hearing and his heart, and setteth on his sight a covering?

Qur'anic Verse - 45:23

2.3.4 Clarifying the meaning of an indefinite word

فَتَلَقَّىٰ آدَمُ مِن رَّبِّهِ كَلِمَاتٍ فَتَابَ عَلَيْهِ
Then Adam received from his Lord words (of revelation), and He relented toward him

Qur'anic Verse - 2:37

The above verse 2:37 talks about indefinite words Adam has received from his Lord. The related verse 7:23 below comes to reveal what those indefinite words were.

قَالَا رَبَّنَا ظَلَمْنَا أَنفُسَنَا وَإِن لَّمْ تَغْفِرْ لَنَا وَتَرْحَمْنَا لَنَكُونَنَّ مِنَ الْخَاسِرِينَ
They said: Our Lord! We have wronged ourselves. If thou forgive us not and have not mercy on us, surely we are of the lost!

Qur'anic Verse - 7:23

2.3.5 Clarifying an indefinite subject through a relative clause

أُحِلَّتْ لَكُم بَهِيمَةُ الْأَنْعَامِ إِلَّا مَا يُتْلَىٰ عَلَيْكُمْ

Lawful for you are the animals of grazing livestock except for that which is recited to you [in this Qur'an]

Qur'anic Verse - 5:1

So, what are these animals which has been mentioned in other verses of the Qur'an? We can find a list of such unlawful livestock in verse 5:3 below.

حُرِّمَتْ عَلَيْكُمُ الْمَيْتَةُ وَالدَّمُ وَلَحْمُ الْخِنزِيرِ وَمَا أُهِلَّ لِغَيْرِ اللَّهِ بِهِ وَالْمُنْخَنِقَةُ وَالْمَوْقُوذَةُ وَالْمُتَرَدِّيَةُ وَالنَّطِيحَةُ وَمَا أَكَلَ السَّبُعُ إِلَّا مَا ذَكَّيْتُمْ وَمَا ذُبِحَ عَلَى النُّصُبِ وَأَن تَسْتَقْسِمُوا بِالْأَزْلَامِ ۚ

Prohibited to you are dead animals, blood, the flesh of swine, and that which has been dedicated to other than Allah , and [those animals] killed by strangling or by a violent blow or by a head-long fall or by the goring of horns, and those from which a wild animal has eaten, except what you [are able to] slaughter [before its death], and those which are sacrificed on stone altars, and [prohibited is] that you seek decision through divining arrows.

Qur'anic Verse - 5:3

2.3.6 Overriding Linguistic Defaults

Often one cannot draw a conclusion from the linguistic meaning of one single verse, without considering the world knowledge and relating to other verses.

Consider for example, verse 6:152 below. Relying only on linguistic information in this single verse results in thinking that when the orphan reaches maturity then it is legal to take from his/her property.

وَلَا تَقْرَبُوا مَالَ الْيَتِيمِ إِلَّا بِالَّتِي هِيَ أَحْسَنُ حَتَّىٰ يَبْلُغَ أَشُدَّهُ ۖ

And do not approach the orphan's property except in a way that is best until he reaches maturity.

Qur'anic Verse - 6:152

However, when we link this verse with verse 4:6 below, we will know that after reaching maturity the property need to be returned to the orphan.

وَابْتَلُوا الْيَتَامَىٰ حَتَّىٰ إِذَا بَلَغُوا النِّكَاحَ فَإِنْ آنَسْتُم مِّنْهُمْ رُشْدًا فَادْفَعُوا إِلَيْهِمْ أَمْوَالَهُمْ

And test the orphans [in their abilities] until they reach marriageable age. Then if you perceive in them sound judgment, release their property to them.

Qur'anic Verse – 4:6

2.3.7 Cause of an action in different Verse

Another type of linkage between verses is the cause relation. That means, in one verse there is an action, and in another verse there is a cause for that action. Let us again see an example.

وَقَالُوا لَوْلَا أُنزِلَ عَلَيْهِ مَلَكٌ ۖ
And they say, "Why was there not sent down to him an angel?"

Qur'anic Verse – 6:8

The above verse did not specify why they want an angel to be sent down. But that reason is mentioned in the verse 25:7 below.

وَقَالُوا مَالِ هَٰذَا الرَّسُولِ يَأْكُلُ الطَّعَامَ وَيَمْشِي فِي الْأَسْوَاقِ ۙ لَوْلَا أُنزِلَ إِلَيْهِ مَلَكٌ فَيَكُونَ مَعَهُ نَذِيرًا
And they say, "What is this messenger that eats food and walks in the markets? Why was there not sent down to him an angel so he would be with him a warner?

Qur'anic Verse - 25:7

2.3.8 Reason of an action in different Verse

ثُمَّ قَسَتْ قُلُوبُكُم مِّن بَعْدِ ذَٰلِكَ فَهِيَ كَالْحِجَارَةِ أَوْ أَشَدُّ قَسْوَةً ۚ
Then your hearts became hardened after that, being like stones or even harder.

Qur'anic Verse – 2:74

The above verse did not specify the reason for their heart being hardened. When bringing the two related verses 5:13 and 57:16 into the picture, we realize the reason being 'breaking the covenant with Allah' in the first, and 'passage of long period' in the second.

فَبِمَا نَقْضِهِم مِّيثَاقَهُمْ لَعَنَّاهُمْ وَجَعَلْنَا قُلُوبَهُمْ قَاسِيَةً
So for their breaking of the covenant We cursed them and made their hearts hard.

Qur'anic Verse – 5:13

وَلَا يَكُونُوا كَالَّذِينَ أُوتُوا الْكِتَابَ مِن قَبْلُ فَطَالَ عَلَيْهِمُ الْأَمَدُ فَقَسَتْ قُلُوبُهُمْ

And let them not be like those who were given the Scripture before, and a long period passed over them, so their hearts hardened

Qur'anic Verse – 57:16

2.3.9 Mentioning the Object of a Subject in different verse

ثُمَّ اتَّخَذْتُمُ الْعِجْلَ مِن بَعْدِهِ وَأَنتُمْ ظَالِمُونَ
Then you took the calf after him, while you were wrongdoers.

Qur'anic Verse – 2:51

The above verse mentions that the children of Israel took a calf. But, they took the calf as what? The object of the verb 'take' was not mentioned in this verse as is the case in many similar verses. However linking these verses with 20:88 makes it clear that they took this calf as their object of worship.

فَأَخْرَجَ لَهُمْ عِجْلًا جَسَدًا لَّهُ خُوَارٌ فَقَالُوا هَٰذَا إِلَٰهُكُمْ وَإِلَٰهُ مُوسَىٰ فَنَسِيَ
And he extracted for them [the statue of] a calf which had a lowing sound, and they said, "This is your god and the god of Moses, but he forgot."

Qur'anic Verse – 20:88

2.3.10 Mentioning the Adverb of Place and Time in another verse

الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ
[All] praise is [due] to Allah , Lord of the worlds

Qur'anic Verse – 1:2

This verse does not qualify the location of this praise. Verse 30:18 comes to qualify this to be throughout the heaves and the earth. And verse 28:70 qualifies the time of this praise to be 'in this life and the Hereafter'.

وَلَهُ الْحَمْدُ فِي السَّمَاوَاتِ وَالْأَرْضِ
And to Him is [due all] praise throughout the heavens and the earth

Qur'anic Verse – 30:18

لَهُ الْحَمْدُ فِي الْأُولَىٰ وَالْآخِرَةِ ۖ
To Him is [due all] praise in the first [life] and the Hereafter.

Qur'anic Verse – 28:70

2.3.11 Specify Semantic Role in a different verse

إِذَا السَّمَاءُ انشَقَّتْ

When the sky has split [open]

Qur'anic Verse – 84:1

فَإِذَا انشَقَّتِ السَّمَاءُ

And when the heaven is split open

Qur'anic Verse – 55:37

وَانشَقَّتِ السَّمَاءُ

And the heaven will split [open]

Qur'anic Verse – 69:16

The three verses above convey the same information i.e., the sky being split on the day of judgement, however, they give no detail of the nature of this splitting but in (25:25) we know that this opening is through emergent clouds. Thus the semantic role of 'medium of split' is defined.

وَيَوْمَ تَشَقَّقُ السَّمَاءُ بِالْغَمَامِ

And [mention] the Day when the heaven will split open with [emerging] clouds,

Qur'anic Verse – 25:25

2.3.12 Clarify a difficult vocabulary in another verse

فَجَعَلْنَا عَالِيَهَا سَافِلَهَا وَأَمْطَرْنَا عَلَيْهِمْ حِجَارَةً مِّن سِجِّيلٍ

And We made the highest part [of the city] its lowest and rained upon them stones of ***sijjil***

Qur'anic Verse - 15:74

In the above verse, the underlined word 'sijjil سجيل' is a rare word and its meaning would be clear when linking this verse with (51:33) which narrates the same story but uses very common word 'طين'/*teen* meaning ‘clay’.

لِنُرْسِلَ عَلَيْهِمْ حِجَارَةً مِّن طِينٍ
To send down upon them stones of clay,

Qur’anic Verse – 51:33

2.3.13 Specifying whether a condition is being fulfilled

Another type of verse relatedness in the Qur’an is when one verse informs about a condition or prohibition is imposed on a past nation or person, and another related verse comes to give the outcome if that nation respected this condition. For example, verse 4:154 prohibits the Children of Israel to transgress on Sabbath, and verse 2:65 informs that these people did transgress afterwards and disobeyed this command or condition imposed on them.

وَقُلْنَا لَهُمْ لَا تَعْدُوا فِي السَّبْتِ
and We said to them, "Do not transgress on the sabbath",

Qur’anic Verse – 4:154

وَلَقَدْ عَلِمْتُمُ الَّذِينَ اعْتَدَوْا مِنكُمْ فِي السَّبْتِ فَقُلْنَا لَهُمْ كُونُوا قِرَدَةً خَاسِئِينَ
And you had already known about those who transgressed among you concerning the sabbath, and We said to them, "Be apes, despised."

Qur’anic Verse – 2:65

2.3.14 Explicit reference to another verse

وَعَلَى الَّذِينَ هَادُوا حَرَّمْنَا مَا قَصَصْنَا عَلَيْكَ مِن قَبْلُ ۖ
And to those who are Jews We have prohibited that which We related to you before.

Qur’anic Verse – 16:118

The above verse explicitly referring to another location through the phrase “which We related to you before”. And when we search for this reference we find that in verse 6:146 this prohibition is mentioned

وَعَلَى الَّذِينَ هَادُوا حَرَّمْنَا كُلَّ ذِي ظُفُرٍ ۖ وَمِنَ الْبَقَرِ وَالْغَنَمِ حَرَّمْنَا عَلَيْهِمْ شُحُومَهُمَا إِلَّا مَا حَمَلَتْ ظُهُورُهُمَا أَوِ الْحَوَايَا أَوْ مَا اخْتَلَطَ بِعَظْمٍ

And to those who are Jews We prohibited every animal of uncloven hoof; and of the cattle and the sheep We prohibited to them their fat, except what adheres to their backs or the entrails or what is joined with bone.

Qur'anic Verse – 6:146

2.4 The Qur'an Translations

As a very influential book, The Qur'an has been translated in most of the living languages, especially languages spoken by majority Muslims. The Qur'an being inimitable words of God, and having unique rhetorical and linguistic style, many translators preferred to translate the 'meaning' of the Qur'an rather than translating the 'Qur'an' itself. (Abdur-Raof 2001, pp. xiv) makes this point clear:

> "Because of the very linguistic and textual nature of the Qur'an, the only way to convey the intended message to the target language reader is to resort to explanatory translation, i.e., the use of footnotes or commentaries to illuminate specific areas in the source text."

Today many translations can be accessed electronically with the flourish of digital media.

The website: [http://www.Quran.org.uk/articles/ieb_Quran_translators.htm] (last accessed on 9th – July-2012) maintains a bibliographical list of many translations including 41 English translations.

2.5 Qur'an Exegesis (Tafsir books)

2.5.1 Significance of books of *Tafsir*

Qur'an exegesis is known in Arabic as books of *Tafsir*. Literally it means to 'interpret'. There has been a huge literature authored by early scholars on this subject. Although the Qur'an was revealed in classical Arabic language, yet *Tafsir* is necessary to understand the Qur'an properly and relate multiple verses when addressing an issue. In section 2.3 I gave several examples where just looking into one verse creates ambiguity, which can only be resolved when relating this verse to others. *Tafsir* books usually discuss such topics and clarify these kind of ambiguity.

Historically, it is known that the Qur'an was not revealed in one shot, rather gradually over a time period of 23 years. During this time, verses of the Qur'an were revealed to Prophet Muhammad based on events and progress of the new religion in Arabia. *Tafsir* books also describe the context of revelation in relation to events in Mecca or Medina.

The rulings of Islam took its final form upon the death of Prophet Muhammad. However, during his messengerhood rulings were evolving gradually towards the final form. For example, as Arabs were used to alcoholic drinks, Islam banned alcohol through a gradual process. Verses 2:219, 4:43 and 5:90 were revealed in that order over a period of five or more years.

يَسْأَلُونَكَ عَنِ الْخَمْرِ وَالْمَيْسِرِ ۖ قُلْ فِيهِمَا إِثْمٌ كَبِيرٌ وَمَنَافِعُ لِلنَّاسِ وَإِثْمُهُمَا أَكْبَرُ مِن نَّفْعِهِمَا ۗ

They ask you about wine and gambling. Say, "In them is great sin and [yet, some] benefit for people. But their sin is greater than their benefit."

Qur'anic Verse – 2:219

يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تَقْرَبُوا الصَّلَاةَ وَأَنتُمْ سُكَارَىٰ حَتَّىٰ تَعْلَمُوا مَا تَقُولُونَ

O you who have believed, do not approach prayer while you are intoxicated until you know what you are saying

Qur'anic Verse – 4:43

يَا أَيُّهَا الَّذِينَ آمَنُوا إِنَّمَا الْخَمْرُ وَالْمَيْسِرُ وَالْأَنصَابُ وَالْأَزْلَامُ رِجْسٌ مِّنْ عَمَلِ الشَّيْطَانِ فَاجْتَنِبُوهُ لَعَلَّكُمْ تُفْلِحُونَ

O you who have believed, indeed, intoxicants, gambling, [sacrificing on] stone alters [to other than Allah], and divining arrows are but defilement from the work of Satan, so avoid it that you may be successful.

Qur'anic Verse – 5:90

The first verse hinted at the harmful side of wine without prohibiting it. The second verse started to prohibit wine in particular time – i.e., before going to prayer – and given that Muslims have to pray five times daily this partial prohibition gradually prepared the Muslims for the final prohibition which came in the final verse. A reader might get confused and would see apparent contradictions when he/she first encounters these verses without knowing the details of this gradual prohibition from books of *Tafsir*.

Hadith – i.e., sayings of the Prophet Muhammad – play a vital role in explaining further Qur'anic verses. Verse 16:44 shows that although the Qur'an was

revealed first to Arabs of Makkah in the Arabic language, yet Prophet Muhammad has to make clear the message of the Qur'an.

وَأَنزَلْنَا إِلَيْكَ الذِّكْرَ لِتُبَيِّنَ لِلنَّاسِ مَا نُزِّلَ إِلَيْهِمْ وَلَعَلَّهُمْ يَتَفَكَّرُونَ
And We revealed to you the message (i.e., the Qur'an) that you may make clear to the people what was sent down to them and that they might give thought.

Qur'anic Verse – 16:44

Books of *Tafsir* usually make this connection between Qur'anic verses and commentaries of Prophet Muhammad. Consider for example the following Hadith.

> Narrated Ibn 'Abbas: Allah's Messenger delivered a sermon and said, "O people! You will be gathered before Allah barefooted, naked and not circumcised. Then (quoting Qur'an) he said, "**As We began the first creation, We will repeat it. [That is] a promise binding upon Us. Indeed, We will do it.** [21:104]" (Bukhari, book 65, hadith no. 4669)

In this Hadith Prophet Muhammad clarified further the meaning of the Qur'anic phrase "the first creation", to mean like the child when born, "barefooted, naked and not circumcised".

2.5.2 Tafsir Methodologies

There have been mainly two approaches by Qur'anic exegetes: tradition-based and opinion-based. The tradition based approach tries to interpret Qur'anic text from traditional sources like, other Qur'anic verses, Hadith, sayings of the companion of Prophet Muhammad or classical Arabic literature. Among the famous books of Tafsir that follow this approach are: *Jami' al-Bayan* authored by *Ibn Jarir at-Tabari* (died in 923 CE), and *Tafsir al-Qur'an al-Adheem* by *Ibn Katheer* (died in 1373 CE). Ibn Katheer summarizes this approach at the introduction of his book (Ibn Katheer d. 1373) vol. 1, pp. 7-9:

> *If a person asks: what is the best way to interpret the Qur'an? The answer is that, the most authentic way is to interpret the Qur'an through the Qur'an, so when a topic is summarized in a place, it will be elaborated in another place. If you still find difficulty, then search the tradition of the Prophet, because Hadith is elaboration of the Qur'an…. Then if we could not find the interpretation in both Qur'an or Hadith, then we return to the sayings of the companions of the Prophet, because they were the most knowledgeable about the Qur'an. They witnessed the context and situation when Qur'an was revealed, and they were possessing complete understanding, authentic knowledge and associated good deeds…. If you*

> *could not find the tafsir in any of the above three, then many of the scholars returned to the sayings of the students of the companions like Mujahid ibn Jabr..and Saeed ibn Jubair..*

The second approach depends more on applying logic and reasoning backed by resorting to general traditions. Among well-known sources following this approach are: Mafateeh al-Ghaib authored by Fakhr ar-Raji (died in 1209 CE) and al-Jamei' Li Ahkam al-Qur'an authored by al-Qurtubi (died in 1272 CE).

2.6 Summary

This chapter discussed the rationale behind choosing the Qur'an and introduced a number of topics related to the domain investigated by this thesis: i.e., the Qur'an. First, the Qur'anic language – Classical Arabic – is compared with the Modern Standard Arabic. Next, a number of unique linguistic characteristics of the Qur'an were discussed for example, the inimitability of the Qur'an, scattered information on the same subject, verb preposition binding, metaphor and figurative use.

Also, this chapter introduced linguistic background on the linguistic subjects investigated in this thesis, namely the pronominal reference system in Arabic, and the text similarity and relatedness in the Qur'an.

As I resorted to both Qur'anic exegesis and translations in this research, this chapter also introduced the significance of these two topics.

Chapter 3
Literature Review

3.1 Computational Qur'anic Studies

In this section, I review a number of computation experiments done on investigating the stylistics of the Qur'an and classification of the Qur'anic chapters.

3.1.1 Computational Clustering

3.1.1.1 Thabet 2005

Thabet (2005) attempted to perform computational cluster analysis by producing a large matrix of word-counts, whose rows are the 114 chapters and whose columns are 3,672 of the Qur'anic words. This short-list of words was obtained after removing: (1) words that appear only once in the Qur'an, and (2) function words like determiners and prepositions. This decision were made based on her assumption that these function words cannot contribute to determination of relationship among surahs. In addition, a purpose-built stemmer was used to consider morphologically variant words as a single word type. The problem Thabet (2005) faced in her analysis is that the shorter chapters produced very sparse rows of values consisting mainly of zeros due to the absence of most of the 3,672 words in any specific chapter. As a solution to this problem, she analysed only the 24 lengthiest chapters which contain more than 1,000 content words. Even this truncation did not solve the sparseness problem as there were still a large number of low frequency words that do not contribute towards thematic cluster formation. As a solution, she had to apply further truncation keeping only the 500 most frequent words.

With this data at hand, Thabet (2005) applied hierarchical clustering algorithms which tend to measure the distance between two chapters based on the shared words. Thus, chapters that share more content words tend to form their own cluster. Figure 3.1 shows the four clusters among the 24 selected chapters denoted by A, B, C and D. Cluster A mainly contains Medinan chapters (with the exception of chapter no. 16 *al-Naḥl* and no. 39 *al-zumar*), and Cluster B are Meccan chapters. Further Medinan chapters are clustered into groups C and D. Her observations can be summarized by these points:

(i) The word ‘Allāh’ is a common theme in all clusters.

(ii) the word ‘qāla’ and its derivaties are a characteristics of Meccan chapters.

(iii) words ‘āmana’ (believe v.), ‘mūᵓmin’ (believer) and ‘ittaqi’ (fear –Allah) are characteristics of Medinan chapters.

(iv) words ‘āyāt (signs), āya’ (sign) are more frequent in Meccan chapters.

(v) Cluster C within Medinan chapters is “more abundant in the use of narratives and addressing Mohamed to provide evidence of his message to people” , whereas chapters in cluster D “are more concerned with addressing believers about the reward for their righteous conduct”

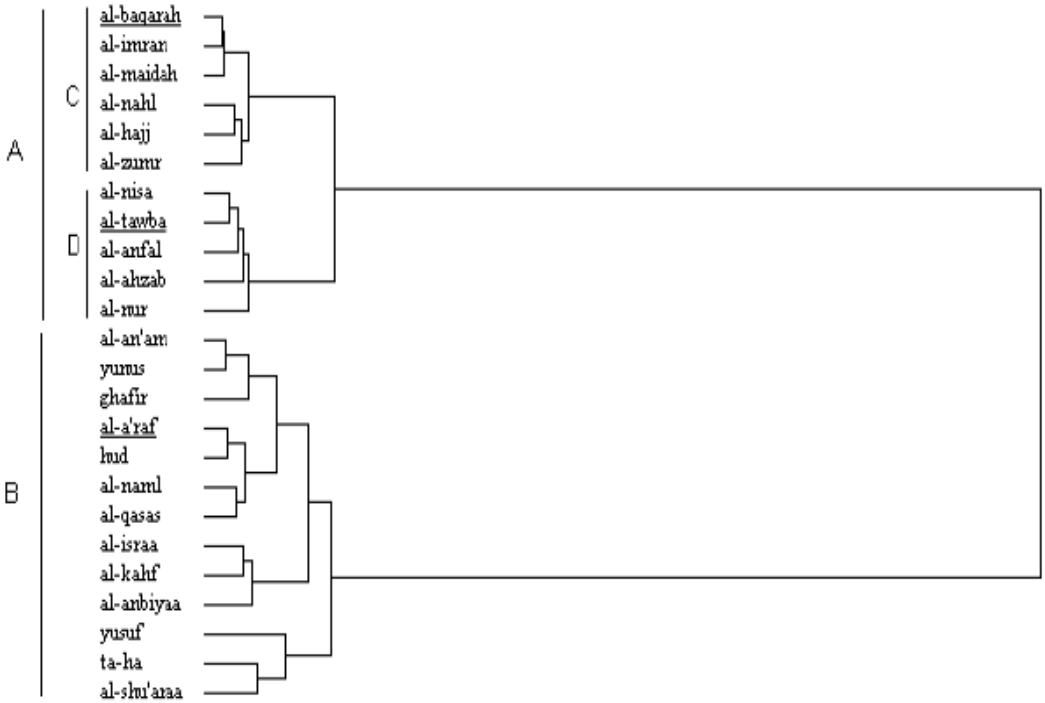


Figure 3.1 – Chapter clusters reported by (Thabet 2005)

Thabet’s work showed a useful application of Machine Learning clustering algorithms, but due to the problem of data sparseness she had to exclude quite a good portion of the Qur’an from her analysis. Chapter 8 describes the machine learning experiments I did to cluster Qur’anic chapters automatically learning from linguistic and domain specific features. Thabet’s statistical measures could have been augmented better with linguistic and domain specific knowledge.

3.1.1.2 Moisl 2009

Moisl (2009) wanted to tackle the problem of shorter chapters that Thabet faced and which resulted in her analysis being limited to only part of the Qur’an. Moisl resorted to sophisticated sampling distribution principles from statistics. He came

up with a minimum value for chapter length. He found that there should be a balance between chapter length and the number of words to be included in cluster analysis, because otherwise considering the entire Qur'an for this analysis is impossible. As a compromise he suggested considering only those chapters which have a word-length of 300 or more. Furthermore, he only considered 9 words as a basis of analysis: Allāh, lā (no), rabb (lord), qāla (said), kāna (was), yaūm (day), nās (mankind), yaūmaʾdh (then), and shar (evil). His clustering results are shown in figure 3.2.

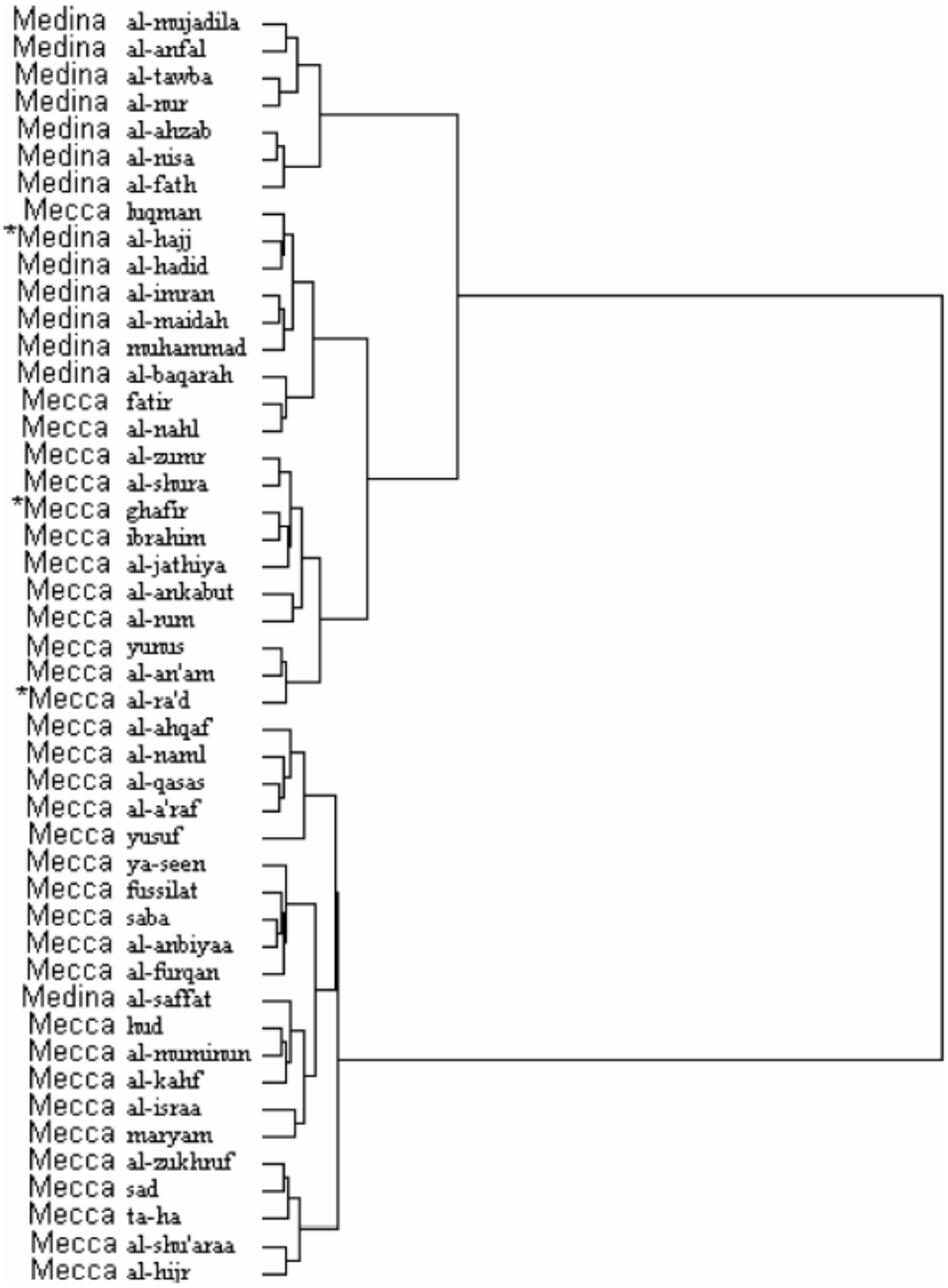


Figure 3.2 – Chapter clusters reported by (Moisl 2009)

However, this analysis is purely based on statistical measures which did not include the entire Qur'an. Moreover, the selection of these 9 words has no overall thematic significance. In my machine learning experiments in chapter 8 Qur'anic chapters were automatically clustered after learning from linguistic and domain specific features. Moisl's statistical measures could have been augmented better with linguistic and domain specific knowledge.

3.1.2 Computational Stylistic Studies

Sadeghi (2011) employed stylometric measures to study the chronology of the Qur'an. He attempted to apply "criterion of concurrent smoothness" to corroborate a seven-phase chronology of the Qur'anic chapters by finding four independent stylistic markers smoothed over this seven phases. His markers were: 1) average verse lengths 2) 28 most common morphemes in the Qur'an 3) frequencies of 114 other morphemes and finally 4) 3693 uncommon morphemes.

Sadeghi thus used computational tools to populate his stylistic markers and enabled him to study chronology of the Qur'an beyond the traditional Meccan and Medinan classification. Others previously attempted the same root, however lack of computational tools stood as obstacles to claim any corroboration. In figure 3.3 below, Sadeghi positions his method (called Modified Bazargan) with those attempted before, namely (Bazargan 1976) and (Weil 1895)

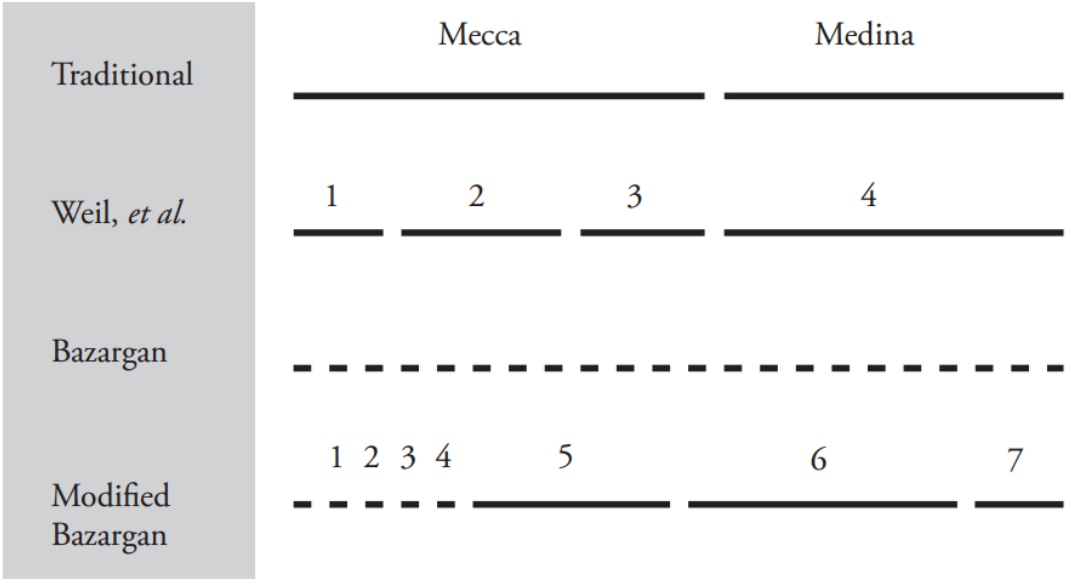


Figure 3.3 – A number of Qur'anic chapter chronology systems reported by (Sadeghi 2011)

The main study engine of Sadeghi has been Multivariate methods employed from statistics aided by computational linguistics analysis tools. Among other findings, Sadeghi concluded from the smoothness of his chosen stylistics markers over the said seven-phases that Qur'an has a single author.

Sayoud (2012) again used authorship discrimination techniques from stylistics to compare parallel texts from Qur'an and Hadith to find out that each maintain distinctive characteristics that entails different authors, as opposed to some claims that Qur'an has been authored by Muhammad. Among many features, Sayoud used are: word frequency, COST parameters, word and character length, number and animal citations and special ending bigrams. COST parameter for a sentence computes the sentence termination similarity (i.e., the terminating syllable) with neighbouring sentence and is used usually in poems.

The markers Sadeghi and Sayoud used were influenced more by traditional stylistic approaches, and thus less attention were given to specific linguistic and domain features of the Qur'an. As is shown in chapter 8, my machine learning experiment heavily relied on such specific features discussed by the Qur'anic scholars.

3.2 Anaphora Resolution

Linguists have debates on the type of knowledge source required to resolve the antecedent of an anaphor. Barbu (2003) cited the following two examples 3.1 and 3.2, to portray the necessity of extra-linguistic world knowledge to reach the intended meaning.

> [3.1] "After considering his dossier, the directory fired the worker because he was a convinced communist".
>
> [3.2] "If a bomb falls next to you, don't lose your head. Put it in a bucket and cover it with sand."

The pronoun 'he' in 3.1 above could equally attach with 'the director' or 'worker', and the only way to resolve is to know if this incidence is taking place in USA or in communist country, for example. In 3.2, the pronoun 'it' could also equally refer to 'the bomb' or 'the head', however considering semantics the 'bomb' is chosen.

Most automatic resolution systems cannot handle these extreme cases, as they lack deep language understanding and linkage with world knowledge. Instead, automatic resolution systems resort to a set of morphological and syntactic rules that covers a great number of cases. For example, SHRDLU (Winograd 1972) used the preference of noun phrase in the subject position as a resolution

method, among others. Hobbs (1978) uses left-to-right breadth-first search on syntactic trees for pronoun resolution.

Also, the theory of Government and Binding developed by (Chomsky 1980) imposes certain constrains on the antecedents a pronoun can refer to, and can be translated into rules in the automatic resolution systems.

Although morphological and syntactic rules caters for a number of anaphor resolution cases, yet for practical situations, the role of semantic analysis still remains a bottleneck. Consider for example a classical example 3.3 below.

[3.3] John took the cake from the table and ate it.

The pronoun ‘it’ syntactically could bind with both ‘cake’ and ‘table’. As deep semantic analysis is a very expensive and difficult ambition, few cheaper attempts have been made. For example, Dagan and Itai, (1990) considered using collocation patterns from a large corpus to model semantic preferences, thus preferring ‘cake’ over ‘table’ as a collocate for the verb ‘eat’ in example [3.3] above.

Effective pronoun resolution system must go beyond the local context and consider discourse analysis. To this end, a number of theories were helpful for resolution systems. For example, Centering Theory (Grosz et at. 1995) models the way attention shifts within discourse units in order to maintain local coherence, and as such, gives a hint regarding where in the discourse a pronoun is likely to find its antecedence.

3.2.1 Automatic resolution systems

Depending on the nature of anaphora and the theory governing it, a number of systems for automatic anaphora resolution exist.

(Hobbs 1978) is an syntax based algorithm which searches syntactic trees for anaphora resolution in a left-to-right, breadth-first manner. Lappin and Leass (1994) used a variety of intra-sentential syntactic factors to construct co-reference equivalence classes. (Brennan et al. 1987) relies on centering theory in pronoun resolution such that the text is coherent.

SHRDLU (Winograd 1972) incorporates one of the earliest anaphora resolvers and relies on knowledge-poor rule-based approach. CogNIAC (Baldwin 1997) is

another rule-based anaphora resolver. Kennedy and Boguraev (1996) modified (Lappin and Leass 1994) by employing heuristic based syntactic rules, thus avoiding deep syntactic parsing. Mitkov et al. (1998) assigns a score to each pronoun antecedent based on a set of boosting and impeding indicators, where at the end the highest score is selected.

Aone and Bennett (1994) developed supervised machine learning algorithm for pronoun resolution. They defined a set of 66 –mostly domain specific- feature sets for business joint venture texts and used them to learn decision trees. Orasan et al. (2000) used genetic algorithms for finding the best combination of weights for the indicators used within the implementation of (Mitkov et al 1998).

Ge et al. (1998) used probabilistic model for anaphora resolution. Their model studied the relative position of antecedents from annotated corpus of Wall Street Journal in terms of syntax and morphological features of the noun phrase. Their approach showed an accuracy of 85%.

3.3 Computational text similarity and relatedness

3.3.1 Evaluation Corpora for Text Similarity and Relatedness experiments

There have been a number of data sources used previously for computational analysis of paraphrased texts and sentence similarity. The Multiple-Translation Chinese Corpus (Huang et at 2002) contains 11 English translations of 105 news stories in Mandarin Chinese, which amounts to 993 sentences. Li et al. (2006) produced a dataset comprising 65 pairs of sentences with similarity scores from 32 human judges. Microsoft Research released a corpus of paraphrased text containing 5801 pairs of sentences collected from various news sources (Dolan et al. 2004). Among this set 3,900 pairs are considered “semantically equivalent” by human judges. Each pair has been visited by at least two annotators with an average agreement rate of 83%.

As for similarity and relatedness at word level where human judges provided scores we can cite: Rubenstein & Goodenough (1965) provided a dataset of 65 word pairs, Miller & Charles (1991) provided a dataset of 30 word-pairs, and

(Finkelstein et al., 2002) with 353 word pairs, and Yang and Powers (2006) created a dataset of 130 verb pairs.

Experiments in word relatedness often use datasets from Scholastic Aptitude Tests (SAT) questions where the most related word pairs need to be selected from five candidates, for example, Jarmasz and Szpakowicz (2003) collected 300 word choice problems.

3.3.2 Knowledge Sources for similarity and relatedness experiments

Dictionaries and thesauri were used in many experiments as well. Popular among them are Roget's Thesaurus (Roget 1962) and Macquarie Thesaurus (Bernard 1986). The most popular knowledge resource for similarity experiments, however, was Wordnets, especially English Wordnet (Fellbaum 1998). Non-English wordnets are available albeit less developed including the Arabic Wordnet (Black et al. 2006). Wikipedia has also been used recently as a knowledge source (Gabrilovich and Markovitch 2007; Zesch et al. 2007).

3.3.3 Semantic relatedness measures

Measuring semantic similarity and relatedness has been an active research topic for a number of years. As far as the knowledge source compilation is concerned, there has been either dependence on linguistic sources like WordNet, or sources constructed by crowdsourcing, like Wikipedia. Zesch and Gurevych (2009) gives a comparative study of both methods.

3.3.3.1 Path-Based Measures

Rada et al. (1989) used path length between two nodes to compute semantic relatedness. Jarmasz and Szpakowicz (2003) adapted (Rada et al 1989) measure to Roget's thesaurus (Roget 1962), while Hirst and St-Onge (1998) adapted Reda measure on WordNet. Leacock and Chodorow (1998) continued the same method but normalized the path length with the depth of the graph. It was noticed that this path-based approach usually faces difficulty when going up in the hierarchy where concepts become very abstract, and thus affects similarity values. Wu and Palmer (1994) introduce the concept of lowest common subsumer as the first shared concept on the path that combines the two concepts.

3.3.3.2 Information Content based measures

Another thread of research on semantic similarity between two words are made surrounding the concept of 'information content', where the similarity or relatedness is measured by the information shared between the two words. Resnik (1995) defines semantic similarity between two nodes as the information content value of their lowest common subsumer. This method relies on calculating the frequency probability of encountering an instance of a word in a large corpus. Jiang and Conrath (1997) modifies Resnik by incorporating information content of the two concepts instead of only the lowest common subsumer. Lin (1998) defines a universal measure derived from information theory.

3.3.3.3 Gloss based measure

Lesk (1986) first introduced a measure of similarity and relatedness between two words based on the shared words in the concept gloss of these words in a dictionary. Banerjee and Pedersen (2002) widens this measure to incorporate the glosses of related concepts. Mihalcea and Moldovan (1999) extended this approach further by taking each noun or verb sense and concatenating the nouns found in the glosses of all WordNet synsets in the sub-hierarchy of this sense.

3.3.3.4 Concept vector based measures

Patwardhan and Pedersen (2006) defined a relatedness measure by computing the cosine angle between the second-order relatedness vector, which was constructed from nouns appearing in the gloss of the two initial concepts. Gabrilovich and Markovitch (2007) define a concept as a measure of the frequency occurrence of this concept in Wikipedia articles, and from that they derived vectors of concepts, and measured relatedness based on cosine similarity of these two vectors. Zesch and Gurevych (2009) describe how they adapted path based and information content based measures to compute semantic relatedness based on Wikipedia category graph and article redirects.

3.3.4 Text similarity and relatedness applications

Semantic similarity and relatedness measures have been used in a number of NLP tasks, like word sense disambiguation (Patwardhan, Banerjee and Pedersen

2003), real word spelling error detection (Budanitsky and Hirst 2006), dialog summarization (Gurevych and Strube 2004).

3.4 Summary

This chapter provides a literature review of some computational works carried out on the Qur'an. For example, Thabet (2005) and Moisl (2009) worked on forming clusters of Qur'anic chapters based on statistical distribution of some keywords. Sadeghi (2011) studied the chronology of Qur'anic chapters by employing stylometric measures.

Major anaphora resolution systems were also introduced in this chapter. Finally, computational text similarity and relatedness was covered by reviewing major evaluation corpora and relatedness measures.

Chapter 4
Linguistic Annotations of the Qur'an

In this chapter I investigate existing annotations of the Qur'an. I start looking into the type of information encoded within the raw text of the Qur'an (like pause mark annotation). Then, I gradually look into existing corpora of the Qur'an with morphological and syntactic tagging. Next, I investigate a number of proposed and work-in-progress efforts to annotate the Qur'an with various linguistic information.

4.1 Raw Qur'anic Text

The raw text of the Qur'an contains certain annotation information that could be a good starting point for a number of corpus linguistic tasks. The Qur'an has been written down shortly after the death of Prophet Muhammad by Caliph Abu Bakr (around 634 CE). Later (around 650 CE) Caliph Uthman commissioned a committee of Prophet's companions to develop a written codex system and make copies to various Islamic regions. Today, we have Muslims all over the world reading the Qur'an in this written form.

The raw text of the Qur'an contains separators for 114 chapters, and each chapter contains separators for verses, and each verse contains a number of pause marks. The smallest unit of text is a verse, and can be referenced by chapter number (or name) and verse number.

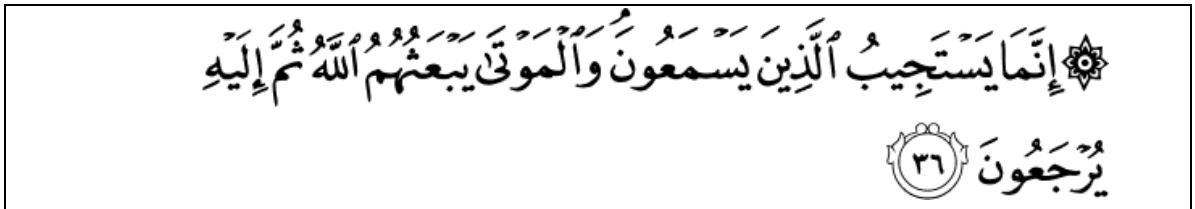
4.1.1 Qur'anic Pause Marks

A Qur'anic verse may contain smaller segments separated by semantically motivated pause marks. These pause marks are added to original codex by scholars for better understanding. To indicate later injection of these marks, they are usually written in smaller superscript codes. Usually any copy of the Qur'an contains as appendix a list of pause marks with explanation of their usage. The following subsection is a rendering of a typical usage from the Qur'an printed by King Fahd Qur'an printing complex. Note that when citing from the Qur'anic verses in this subsection, scanned images of Arabic verses are included to portray the Arabic pause marks used in original Qur'an text, also English translations are modified to make the semantic significance of the pause marks clearer. Also, note that none of the English translations I came across incorporates these pause marks as it appears in the Arabic version, rather they

have used the traditional English punctuation marks for the translator's convenience and understanding for the target translation language.

4.1.1.1 Mandatory Pause (M)

Consider verse 6:36 below. The small Arabic letter (م) (M) is placed to impose a mandatory pause in the sentence after the phrase 'Only those can accept who hear'. If this pause is not observed, then the meaning will change and would mean that, the dead will also accept.



Literal:

Only accepts those who hear (M) and the dead Allah will raise them …

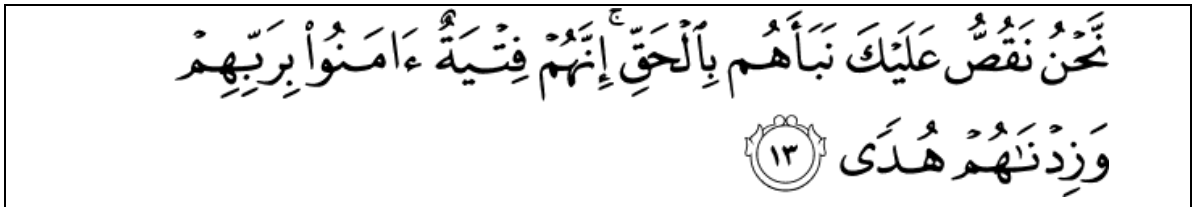
Pickthall:

Only those can accept who hear **(M)** As for the dead, Allah will raise them up; then unto Him they will be returned.

Qur'anic Verse – 6:36

4.1.1.2 Voluntary Pause (V)

This mark is indicated by small (ج) (V) letter over the place of pause. Consider as an example verse 18:13 below. The pause mark (V) is an optional pause that has no effect on the meaning of the verse if the reader pauses or continues.

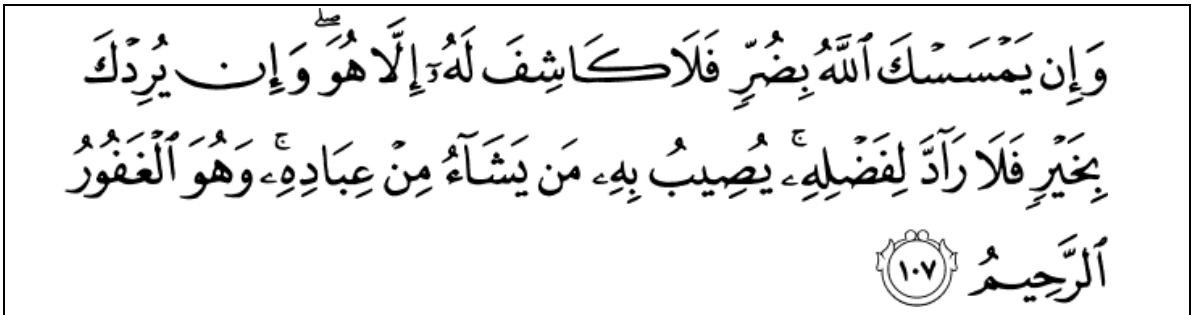


We narrate unto thee their story with truth **(V)** Lo! they were young men who believed in their Lord, and We increased them in guidance.

Qur'anic Verse – 18:13

4.1.1.3 Voluntary Pause where continuation preferred (VC)

This mark is indicated by Arabic superscripted small letters (صلى) (VC) over the place of pause. A reader may pause at this location, however continuation is preferred. Consider for example, verse 10:107 below. The continuation at (VC) is more appropriate as it completes the both part of Allah's supreme power in removing adversity or intending good. However, in the other two (V) locations, meaning has no effect whether the reader pauses or continues.

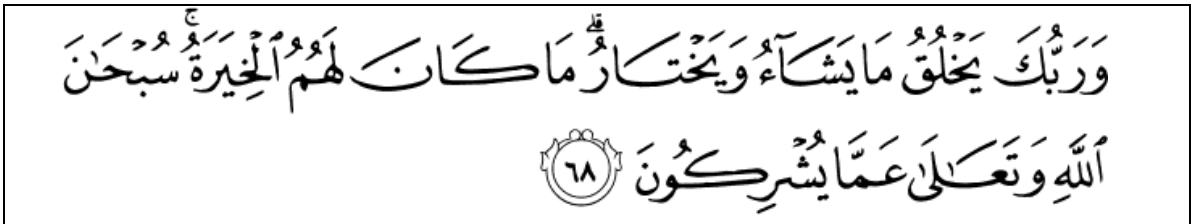


If Allah afflicteth thee with some hurt, there is none who can remove it save Him; **(VC)** and if He desireth good for thee, there is none who can repel His bounty. **(V)** He striketh with it whom He will of his bondmen. **(V)** He is the Forgiving, the Merciful.

Qur'anic Verse – 10:107

4.1.1.4 Voluntary Pause where pause preferred (VP)

This marker is opposite of VC and is indicated by small letters (قلي) (VP) over the pause location. A reader may continue reading at this location, however pausing is more appropriate considering the meaning of the verse. Consider for example verse 28:68 below, pausing over (VP) emphasis the supreme power of Lord in creating and choosing, and distinguished that from the choice of other creatures.

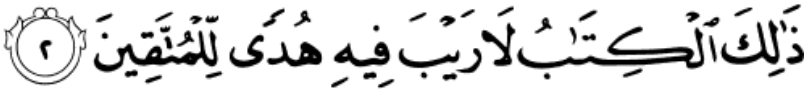


Thy Lord bringeth to pass what He willeth and chooseth **(VP)**. They have never any choice.**(V)** Glorified be Allah and Exalted above all that they associate (with Him)!

Qur'anic Verse – 28:68

4.1.1.5 Exclusive OR pause (XOR)

This kind of mark is not found often in the Qur'an, and is indicated by a pair of two three-dot triangle shape. For simplicity I denoted that as XOR. The reader has option to pause in one of these pairs, however, when pausing in one, he/she must continue in the other. Each instance of reading gives a meaning different thaى the other instance. Consider for example, verse 2:2 below.



This is the Book no doubt **(XOR)** in it **(XOR)** a guidance for those conscious of Allah.

Qur'anic Verse – 2:2 – Translation modified from Pickthall to express the pause mark significance

The reader have two options to read this verse:

- When pausing in the second marker, this will be the reading: *This is the Book no doubt in it. A guidance for those conscious of Allah.*
- When pausing in the first marker, the reading will be: *This is the Book no doubt. In it a guidance for those conscious of Allah.*

Note that most English translations I came across did not take this pause mark into consideration and mostly assumed the meaning when pausing on the second marker.

4.1.2 Qur'anic Pause Mark Annotation

(Brierley, Sawalha and Atwell 2012) has developed a coarse-grained boundary annotation scheme incorporating the above mentioned pause markers. They have collapsed eight-degree boundary strengths to the three major boundary types: {major, minor, none}. As a result they ended up with 8,230 sentences of the Qur'an marked with these three markers. A probabilistic tagger was trained using this corpus to predict these boundaries with a success rate of 86.62% (Sawalha, Brierley and Atwell 2012).

4.2 Morphological Annotation

4.2.1 Haifa Corpus

The first morphology annotation of the Qur'an for computational research was done –to the best of our knowledge- by Haifa University team (Dror et al. 2004).

The authors used the Xerox finite-state toolbox (Beesley and Karttunen 2003) which was fed with lexicon and rules to generate morphological annotation.

4.2.1.1 Lexicon

Haifa corpus collected Qur'anic lexicon from an existing concordance of the Qur'an (Abdulbaqi 1955) and manually constructed three classes: closed-class functional words, nominal bases and verbal bases. The lexicon associates with each lexeme its root and pattern, and sometimes more information like surface form, inflected forms (like concatenation of particles, gender and number), and continuation class that specifies which suffixes a lexeme can combine with. For example, consider the following sample entry for a noun lexeme:

```
swr+fu&lat:suurat    NounEndingFem;
```

The surface word is the noun '`suurat`' and the root is '`swr`' which has patter of '`fu&lat`' and continuation class of '`NounEndingFem`' which indicates that the lexeme can be suffixed by feminine nominal affixes.

As for the verbs, the authors collected Qur'anic verbal root list from (Chou´emi 1966; Ambros 1987) and generated all possible instantiations of these roots in all verbal patters of Qur'anic Arabic, which resulted in 100,000 possible verb bases. A major portion of these bases are pruned off benchmarking with only verbal forms that exists in the Qur'an. Following is a sample lexical entry of a verb form that is prefixed with a particle:

```
fa-'akalaa -> fa+Particle+Conjunction+'kl+Verb+Stem1+Perf+Act+3P+Dual+Masc
```

note that the affixed particle '`fa`' is separated from the verb '`'akalaa`' with a dash '-' and each gets separate morphological features.

4.2.1.2 Finite-State Rules

As the lexicon generates the base forms with additional information, Arabic grammar rules are introduced to filter out redundancies, to handle morpho-phonological alteration, to take care of idiosyncrasies –like verbal weak paradigms and to implement purely phonetic rules.

For example, a grammar rule dictates that the noun occurring after a preposition must be in the genitive case. Applying this rule would prune out redundant cases from the lexicon that does not comply with this rule. Thus, pruning is done on the lexicon to reduce the number of possible targeted outcomes.

In total there were approximately 50 rules for nouns and 300 for verbs. Applying these rules reduced the original lexicon entries but still kept for certain entities multiple possibilities, with an average of 1.37 analysis per token.

4.2.1.3 Query Interface

Figure 4.1 is a screenshot of the interface as reported in (Dror et al 2004). Each query enables users to search for various morphological features such as number, gender, case, aspect, etc. The user can request certain root or pattern. Context sensitive options appears for different parts of speech. For example, when selecting verb, the user can check any of the 12 available stems in Arabic verbs.

4.2.1.4 Evaluation

The Haifa corpus was evaluated for accuracy by manually checking system performance for chapter 8 of the Qur'an, which consists of 1248 words (approx. 1.6% of the Qur'an). Their evaluation system produced 1440 analysis with an average of 1.15 outcome for each word. Benchmarking against manual annotation, 69 of these analyses were incorrect, 205 as possible and 1162 correct analysis. The authors report 93% recall, 80% precision and an f-measure of 0.86.

This corpus was first of its kind and was intended for teaching and learning, so Arabic grammar students can search for evidence of certain morphological features within the Qur'an. However, the accuracy level of this corpus needs to be manually verified by scholars to encompass the entire Qur'an and not only 1.6% of the text in order for this corpus to be widely used by the community.

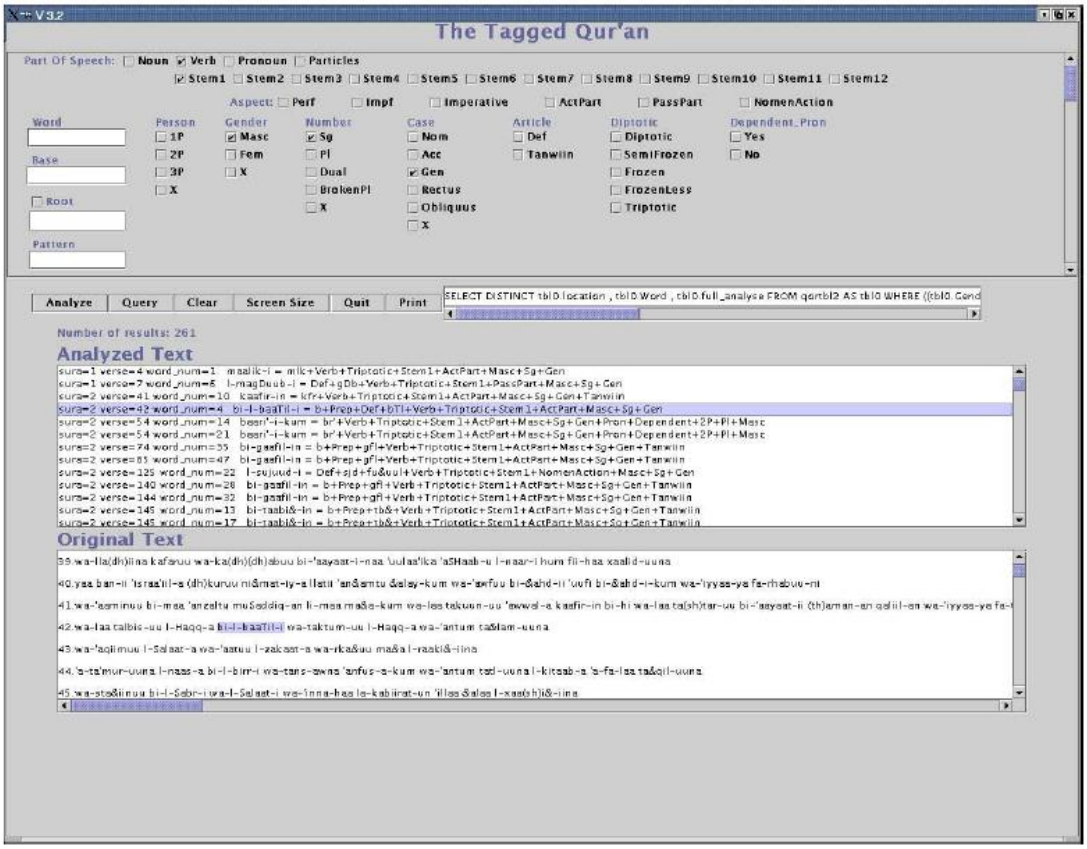


Figure 4.1 – Query Interface for Haifa database (Dror et al 2004)

4.2.2 Qur'anic Arabic Corpus (QAC)

This corpus started by an initial tagging of the Qur'an text by adapting Buckwater Arabic Morphological Analyser (Buckwalter, 2002) to work with Qur'anic text. This automatic annotation was then subjected to manual verification through a collaborative online platform.

Figure 4.2 below shows a sample morphology analysis for word no. 16 in verse 11:28.

Figure 4.2 – Morphology analysis for word 11:28:16 from QAC

Following features are displayed with each Qur'anic word:

- Morphological segments of a Qur'anic word. For example, in the sample I notice that this word has four segments depicted by different colours.
- POS tagging for each segment. For example, in the sample, two pronouns (PRON), a verb (V) and an interrogative (INTG) is shown.
- Expanded narrative on each POS.
- Arabic auto generated translation of the detailed morphological feature.
- Transliteration of the Arabic word
- Word meaning in English.

QAC employed specialized POS tag set to suite the Qur'anic Arabic. For example, the tag INL refers to Qur'anic initials repeated at the beginning of many Qur'anic chapters. Also, the morphological feature set covers a wide range of freatures like: prefix freatues, root, lemma, person, number, aspect (perfect, imperfect, imperative), mood (indicative, subjunctive, jussive, energetic), voice, verb form (from I to XII), derivation (active participle, passive participle, verbal noun), state (definitive, indefinitive), case (nominative, accusative, genitive) and suffix features.

QAC corpus is available for online[1] browsing and downloading for research purposes. QAC contrasts with the Haifa corpus as it integrates the corpus online for browsing and query, and provides a collaborative platform allowing for the continuous refinement and improvement of the annotation. QurAna annotation of pronouns was based on PRON tags from QAC.

4.3 Syntactic Annotation

4.3.1 Qur'anic Arabic Dependency Treebank

Treebanks are parsed corpora with each sentence of including texts are annotated with syntactic information (e.g., sentence head, noun pharase, verb phrase, etc.) in a tree structure. Qur'anic Arabic Dependency Treebank (QADT) (Dukes et al 2010) used traditional Arabic grammar rules to visualize the syntax of Qur'anic verse segments using dependency graphs. These graphs serve two purposes: a) being a useful educational resource for students of Arabic grammar,

[1] Available at http://corpus.Quran.com

and b) as the graphs are machine readable, this will enable future computational research to learn automatic parsers for traditional Arabic grammar.

As was the case with building the morphology in QAC, QADT went into an iterative process to minimize the manual effort to build the Treebank. A rule-based dependency parse was developed specifically for Qur'anic Arabic which created the first base of the Treebank with a reported accuracy of 78% (Dukes and Buckwalter 2010). Next, the automatic annotation was manually corrected resorting to trusted publications on the Qur'anic grammar. Finally, an extensive annotation guideline was produced and the Treebank was uploaded online[1] for further collaborative verification.

The approach followed in QADT contrasts to other Arabic treebanks -such as Penn Arabic Treebank (Maamouri et al 2004) or Prague Arabic Dependency Treebank (Smrz and Hajic 2006)- in incorporating traditional Arabic grammar, while others adapted an existing English scheme for Arabic, causing much of the peculiarities of the language to be lost.

Figure 4.3 is a sample parse tree for a segment from verse 1:1

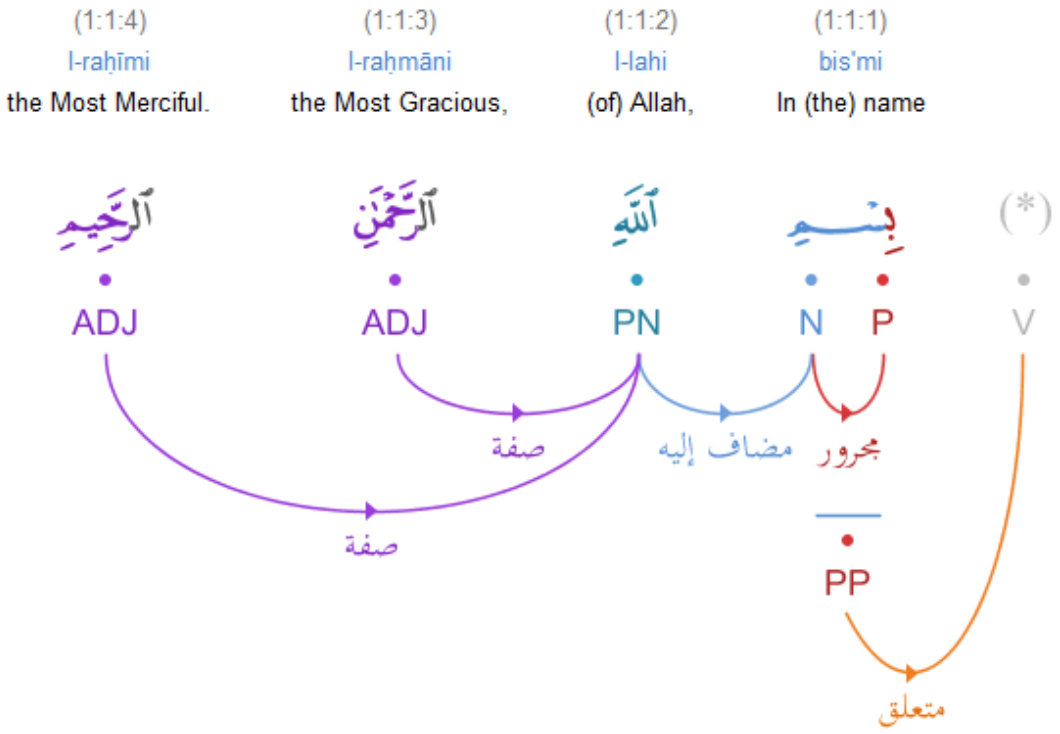


Figure 4.3 – Syntactic annotation for verse 1:1 from QADT

[1] Available at http://corpus.Quran.com

The following features are incorporated within these syntactic graphs:

- Directed arrows shows the dependency direction. For example, both the third and fourth word in verse 1:1 (الرحمن) and (الرحيم) are adjectives describing the second word (الله), hence the direction of the arrows.
- A clause constituted from more than one morpheme can function collectively as a dependency. For example, in the first word of verse 1:1, both the noun and particle collectively act as a prepositional phrase pointing to a verb.
- As Arabic is a pro-drop language, certain grammatical elements could be dropped from a sentence, for example the head verb of a prepositional clause (as in verse 1:1 above, where the dropped verb is placed as asterisk), or hidden pronouns.

QADT is a work in progress and approximately 43% of the Qur'an was covered at the time of writing.

4.3.2 Pronoun antecedent annotation

QurAna is a corpus that tags all pronouns of the Qur'an (approx. 24,000 pronouns). Figure 4.4 gives an example from verse 38:29. All explicit pronouns identified through PRON tag in QAC are tagged with their referent, which could be mentioned earlier or not. In both cases, a concept ontology is maintained.

Verse 10:58

Say: In the bounty of Allah and in His mercy: therein let them rejoice. It is .better than what they hoard	قُل بِفَضلِ اللَّهِ وَبِرَحمَتِهِ فَبِذَٰلِكَ فَليَفرَحوا هُوَ خَيرٌ مِمّا يَجمَعونَ

gloss	Concept	Verse	Pronoun context	Antecedent	#
Allah	الله	10:58	وَرَحْمَةٌ لِّلْمُؤْمِنِينَ قُلْ بِفَضْلِ ٱللَّهِ وَ بِ رَحْمَتِهِۦ فَبِذَٰلِكَ فَلْيَفْرَحُوا۟ هُوَ خَيْرٌ مِّمَّا يَجْمَعُونَ	الله	1
Muslims	المسلمون	10:58	قُلْ بِفَضْلِ ٱللَّهِ وَبِرَحْمَتِهِۦ فَبِذَٰلِكَ فَ لْ يَفْرَحُوا۟ هُوَ خَيْرٌ مِّمَّا يَجْمَعُونَ قُلْ أَرَءَيْتُم		2
the bounty of Allah	فضل الله	10:58	بِفَضْلِ ٱللَّهِ وَبِرَحْمَتِهِۦ فَبِذَٰلِكَ فَلْيَفْرَحُوا۟ هُوَ خَيْرٌ مِّمَّا يَجْمَعُونَ قُلْ أَرَءَيْتُم مَّآ	بِفَضْلِ اللَّهِ وَبِرَحْمَتِهِۦ	3
Muslims	المسلمون	10:58	فَبِذَٰلِكَ فَلْيَفْرَحُوا۟ هُوَ خَيْرٌ مِّمَّا يَجْمَعُونَ قُلْ أَرَءَيْتُم مَّآ أَنزَلَ ٱللَّهُ لَكُم		4

Figure 4.4 - Pronoun resolution of verse 10:58

QurAna is one of the main novel contributions of this research, and chapter 5 is entirely dedicated to describing this corpus.

4.4 Semantic Annotation

4.4.1 Frame Semantics

Frame semantics is introduced by Fillmore (Fillmore 1976). Based on this linguistic theory, researchers in the International Computer Science Institute (ICSI), Berkeley, started the FrameNet project (Ruppenhofer et al 2005) (Baker et al 1998) (Fillmore et al 2003) in 1997 to build an online lexicon for English frames which are to capture the semantic and syntactic properties of English predicates based on their usage in the British National Corpus (BNC) (Aston & Burnard 1998). Based on the experience of the English FrameNet, various projects started to build similar lexicon for other languages.

FrameNet is a lexicon that describes 'Frames' as a schematic representation describing a situation involving various conceptual roles called 'Frame Elements (FE)'. A frame can be 'evoked' by a group of related predicates (mainly verbs, but also nouns or adjectives) called 'Lexical Units (LU)'.

For example, the verb 'buy' along with 'purchase' form the LUs that can evoke the `commerce_buy` frame. This frame has 'core' frame elements that are essential to the meaning of the frame e.g., FEs (BUYER, GOODS) and has many other non-core FEs (like: DURATION, MANNER, MEANS, MONEY, PLACE, PURPOSE, RATE, REASON, RECIPIENT, SELLER, TIME, UNIT).

Following is an example from `commerce_buy` frame description. (The lexical unit is in **boldface** and Frame Elements are in CAPITALS).

`[BUYER Lee]` **`BOUGHT`** `[GOODS a textbook] [SELLER from Abby]`

Currently, the FrameNet project contains more than 10,000 lexical units in nearly 800 hierarchically related semantic frames, exemplified in more than 135,000 annotated sentences. (Ruppenhofer et al 2005).

A preliminary study of the Qur'anic frame semantics was described in (Muhammad and Atwell 2009). A sample Qur'anic verbs were analysed and benchmarked against English FrameNet. For example the verb eat according to the FrameNet definition belongs to `'Ingestion'` frame, which has two core

elements: `ingestible` and an `ingestor`. Looking into the valences of this verb in the Qur'an, I observe that it appeared –with derived forms- 100 times. Table 4.1 below lists a few representative concordance lines. In the majority of the cases, its use was in alignment with FrameNet descriptions, like the example of line [A]. However, there are examples where 'eat' is used differently, for example lines [B] uses 'eat' to mean 'eating money' which is not a usual ingestible item, and hence it means to 'earn money unlawfully'. Consider also the line [E] where 'seven years' are the 'ingestor' which violates the 'sentient' restriction of FrameNet.

A	the sea to be of service that ye	**eat**	fresh meat from thence	16:14
B	And	**eat**	not up your property among	2:188
C	Would one of you love to	**eat**	the flesh of his dead brother?	49:12
D	seven fat kine which seven lean were	**eating**		12:43
E	seven hard years which will	**eat**	all that ye have prepared for them	12:48
F	they	**eat**	into their bellies nothing else than fire	2:177
G		**eat**	of unlawful	5:42

Table 4.1 – Sample usage of the verb 'eat' in the Qur'an

For another example, take the FrameNet's `Giving` frame. Following verses are labelled with the frame elements DONOR, THEME, DONATED_AMOUNT and RECIPIENT.

[A] and [they DONOR] spend out of [what We have provided for them THEME] [2:3]

[B] and they ask you what [they DONOR] should spend. Say, "[the excess DONATED_AMOUNT]". [2:219]

[C] and they ask you what they should spend. Say, "Whatever [you DONOR] spend of [good DONATED_AMOUNT] is [for parents and relatives and orphans and the needy and the traveller RECIPIENT]. [2:215]

While [A] talks about the theme of the donated money, [B] qualifies the type of this theme to be from the excess money that is left after spending on the necessary needs. However, [C] gives an answer to the same question as in [B],

but specifies the recipient of this spend rather than the type or amount of the money.

Similar to verbs, some frequent Qur'anic nouns can also dictate useful semantic frames. As another example, following mind map depicts various thematic slots the noun 'praise' and its derivatives can dictate.

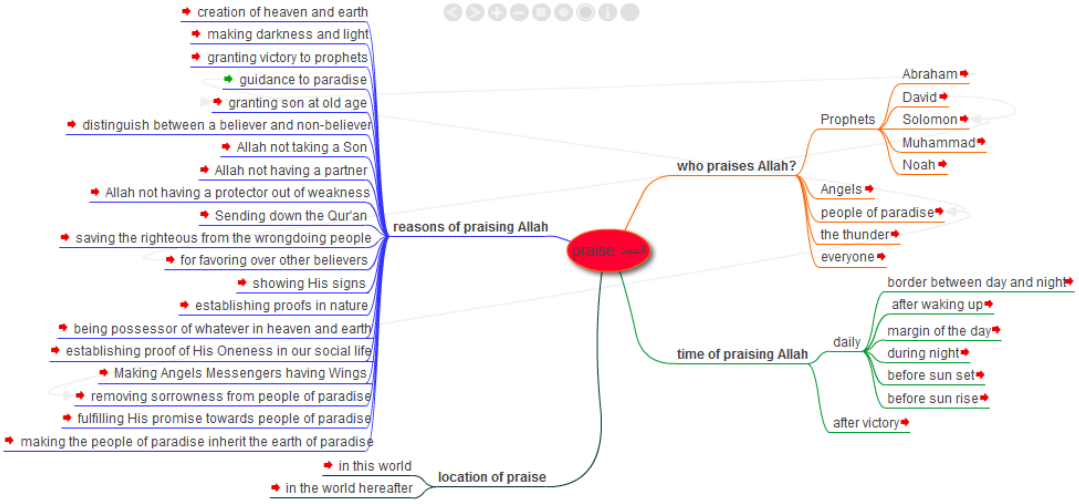


Figure 4.5 – Semantic frames for the concept 'praise' in the Qur'an

All works reviewed so far were proof-of-concept attempts towards semantic frames of the Qur'an. When completed these kind of annotations will be very helpful for knowledge discovery and question-answering from the Qur'an.

4.4.2 Qur'anic Prepositional Verbs

(Fiteih 1983) studied the prepositional verbs considering the Qur'an as his corpus. He could classify four classes of Qur'anic verbs based on the number and type of nominals and prepositions these verbs allow. There are cases when a verb allows one prepositional object (e.g., *reach to something* as in verse 11:70), or a nominal and a prepositional object (e.g., let loose on someone something as in verse 54:19), or two prepositional objects (e.g., come forth unto someone from some place as in verse 19:11), or one nominal object and two prepositional objects as in verse 16:71 or one prepositional object and two nominal objects as in verse 4:95. Please note how the two prepositional objects appear adjacent in the original text in 4:95.

فَلَمَّا رَأَىٰ أَيْدِيَهُمْ لَا **تَصِلُ** إِلَيْهِ نَكِرَهُمْ
And when he saw their hands **reached** not to it, he mistrusted them..

Qur'anic Verse – 11:70

إِنَّا **أَرْسَلْنَا** عَلَيْهِمْ رِيحًا صَرْصَرًا

Lo! We **let loose** on them a raging wind

Qur'anic Verse – 54:19

فَخَرَجَ عَلَىٰ قَوْمِهِ مِنَ الْمِحْرَابِ

Then he **came forth** unto his people from the sanctuary

Qur'anic Verse – 19:11

وَاللَّهُ **فَضَّلَ** بَعْضَكُمْ عَلَىٰ بَعْضٍ فِي الرِّزْقِ

And Allah hath **favoured** some of you above others in provision

Qur'anic Verse – 16:71

وَفَضَّلَ اللَّهُ الْمُجَاهِدِينَ عَلَى الْقَاعِدِينَ أَجْرًا عَظِيمًا

Literal:

And Allah hath **bestowed** on those who strive above the sedentary a great reward.

Pickthall:

He hath **bestowed** on those who strive a great reward above the sedentary

Qur'anic Verse – 4:95

4.5 Other annotation research

4.5.1 Qur'anic Concept Ontology

QAC made an initial attempt to build a concept ontology of the Qur'an based on knowledge extracted from traditional sources like traditions of Prophet Muhammad (i.e., *Hadith*) or books of Qur'anic exegesis (i.e., *Tafsir*). QAC ontology is represented visually online[1] through a network of 300 linked concepts with 350 relations. Figure 4.6 shows an excerpt from the graph containing the concept category 'physical substance'.

[1] Available online at http://corpus.Quran.com/ontology.jsp

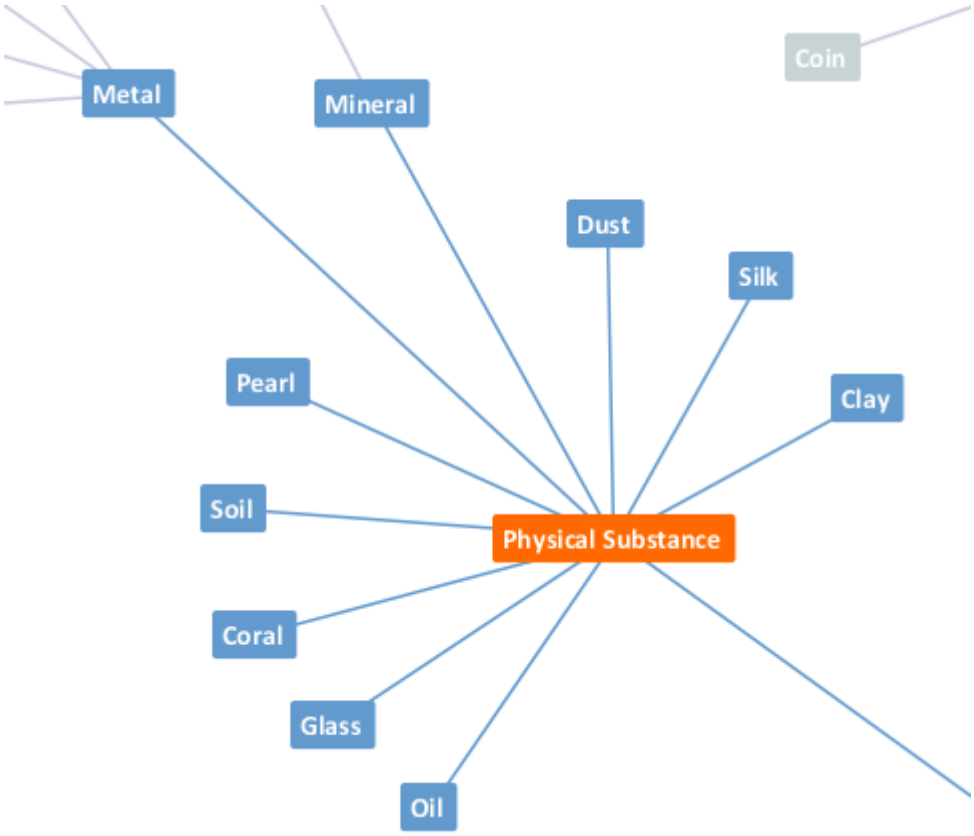


Figure 4.6 - partial concept graph from QAC showing the category 'physical substance'

The graph allows drilling down from the concept category until reaching at the lowest level to the concept, where a number of pieces of information are shown for each concept as follows.

- A sample verse from the Qur'an containing this concept.
- Link to Wikipedia entry of this concept.
- Link to all occurrences of this concept in the Qur'an.
- Visual concept map of neighbouring concepts.
- Predicate logic relations of this concept

4.5.2 Annotation of intonation and pronunciation (*Tajweed*)

The Qur'an has propagated over centuries through oral transmission from Angel Gabriel to Prophet Muhammad to his companions and so on. The written format according to specialized codex was a supportive mechanism to preserve the Qur'an from getting extinct. However, the official form continued to be through oral teaching whereby the student learns the proper pronunciation and intonation sound of each verse.

(Taha 2008) designed a colour-coding mechanism to annotate Qur'anic text with selected rules of intonation, known as *Tajweed* rules. Figure 4.7 is an illustrative verse where this colour coding annotation is employed.

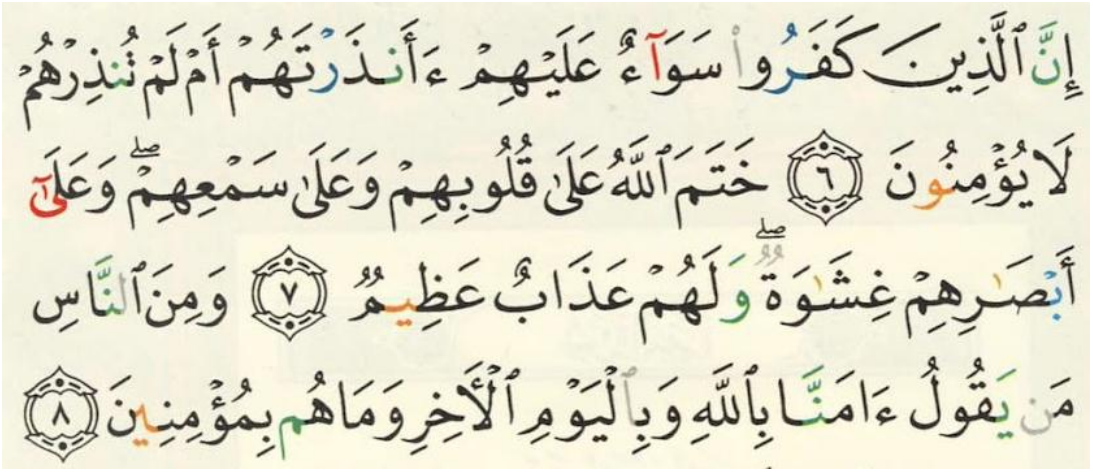


Figure 4.7 – excerpt from (Taha 2008) showing colour coded annotation of tajweed rules

Following are the main coloured features in this annotation:

- The dark red colour indicates necessary prolongation of six vowels each of which is about half a second.
- Blood red colour indicates obligatory prolongation of five vowels.
- Orange red colour indicates permissible prolongation from two to six vowels.
- Cumin red colour indicates normal prolongation of two vowels.
- Green colour indicates nasalization sound accompanying a prolongation of two vowels.
- Grey colour indicates silent letters.
- Dark blue colour indicates emphatic pronunciation of the Arabic letter (ر) R.
- Blue colour indicates echoing sound (qalqalah).

4.6 Qur'an as a Corpus

The size of the Qur'an is relatively small (i.e., approx.. 128k word segments) comparing with standard contemporary corpora. This size might not be large enough for some empirical language learning experiments. However, Qur'an is a unique text that has the significance of being verbatim words of God – as

believed by Muslims worldwide. Moreover, this text -although small in size- has been a subject of language and literary study for over 14 centuries. Given these facts I advocate the Qur'an to be a corpus given the broad sense of this term from pragmatic perspective –as opposed to semantic meaning- as suggested by (Kilgarriff and Grefenstette 2003):

> *A corpus is a collection of texts when considered as an object of language or literary study.*

However, in this section, I would still want to study the qualifications of the Qur'an being a corpus from the specific and modern connotation of the word "corpus" as described in (McEnery and Wilson 1996):

> *In principle, any collection of more than one text can be called a corpus. . . . But the term "corpus" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition provides for. These may be considered under four main headings: sampling and representativeness, finite size, machine-readable form, a standard reference.*

4.6.1 Sampling and Representativeness

The Qur'an is considered to be a sample of Classical Arabic, but also is a unique sample of Words of God. The Qur'an describes in one verse that Words of God is infinite (see verse 31:27 below). Included to this unique corpus would be the Bible as well. However, because: a) the original text of the Bible was not in Arabic, and b) it is difficult to judge which part of the Bible at our hand is verbatim Words of God.

وَلَوْ أَنَّمَا فِي الْأَرْضِ مِن شَجَرَةٍ أَقْلَامٌ وَالْبَحْرُ يَمُدُّهُ مِن بَعْدِهِ سَبْعَةُ أَبْحُرٍ مَّا نَفِدَتْ كَلِمَاتُ اللَّهِ ۗ إِنَّ اللَّهَ عَزِيزٌ حَكِيمٌ
And if all the trees in the earth were pens, and the sea, with seven more seas to help it, (were ink), the words of Allah could not be exhausted. Lo! Allah is Mighty, Wise.

Qur'anic Verse - 31:27

The Qur'an has been used by early Arabic grammarians as supportive evidence of various linguistic phenomena of classical Arabic grammar. The Qur'an itself testifies in multiple locations its significance being revealed in pure and clear Arabic language. Following are some sample verses.

إِنَّا جَعَلْنَاهُ قُرْآنًا عَرَبِيًّا لَّعَلَّكُمْ تَعْقِلُونَ

Lo! We have appointed it a Lecture, in Arabic that haply ye may understand.

Qur'anic Verse - 43:3

وَإِنَّهُ لَتَنزِيلُ رَبِّ الْعَالَمِينَ نَزَلَ بِهِ الرُّوحُ الْأَمِينُ عَلَىٰ قَلْبِكَ لِتَكُونَ مِنَ الْمُنذِرِينَ بِلِسَانٍ عَرَبِيٍّ مُّبِينٍ

And lo! it is a revelation of the Lord of the Worlds, Which the True Spirit hath brought down, Upon thy heart, that thou mayst be (one) of the warners, In plain Arabic speech.

Qur'anic Verse - 26: 192-195

قُرْآنًا عَرَبِيًّا غَيْرَ ذِي عِوَجٍ لَّعَلَّهُمْ يَتَّقُونَ

A Lecture in Arabic, containing no crookedness, that haply they may ward off (evil).

Qur'anic Verse - 39:28

If representativeness is measured by the extent to which the Qur'anic vocabulary includes full-range vocabulary of the classical Arabic language, then –as we will see in the next section- the Qur'an contains only 13.5% of the overall Arabic roots. However, considering the content and the usefulness of the text for society then the Qur'an itself claims to be self-sufficient and containing the key to all solutions. See some supportive verses below.

وَلَا يَأْتُونَكَ بِمَثَلٍ إِلَّا جِئْنَاكَ بِالْحَقِّ وَأَحْسَنَ تَفْسِيرًا

And they bring thee no similitude but We bring thee the Truth (as against it), and better (than their similitude) as argument.

Qur'anic Verse - 25:33

فَإِمَّا يَأْتِيَنَّكُم مِّنِّي هُدًى فَمَنِ اتَّبَعَ هُدَايَ فَلَا يَضِلُّ وَلَا يَشْقَىٰ

But when there come unto you from Me a guidance, then whoso followeth My guidance, he will not go astray nor come to grief.

Qur'anic Verse - 20:123

A quick examination of the Qur'an shows wide coverage of topics scattered in the Qur'an, from eschatological topics, to attributes of God, to stories of ancient people, to matters related to monetary transaction to jurisprudence related to marriage and maternity, to relationship with other religions, to the description and lessons from major battles at the time of Prophet Muhammad. From this perspective, the Qur'an can be considered a general corpus that encompasses a number of domains and genres. Also, the Qur'an could be considered balanced (over the 23 years of Qur'an revelation: approx. between years 609 – 632 CE) because of the distribution of these topics evenly in various locations of the Qur'an. However, linking a topic to Qur'anic text is not always trivial and often deep language, world knowledge and contextual clues are required in order to link the text to the topic. In this regard coupling the Qur'an with Tafsir sources could be beneficial.

4.6.2 Finite Size

The Qur'an is finite in size. Thus, it is a static sample corpus. The Qur'an contains 77,804 space delimited words. However, considering word segments we have 127,795 segments. In terms of word types the Qur'an contains 14,948 word types which can be reduced to 1,619 root words. These figures were captured thanks to the morphological corpus of the Qur'an (Qur'anic Arabic Corpus). To compare with the overall available Arabic root words, the figure approximates to 12,000 roots as enumerated in a reference Arabic lexicon called Taj-ul-A'rous (Al-Zubaidy 1790).

4.6.3 Machine-readable form

The Qur'anic text has been digitized and made into a machine readable format with proper Unicode encoding. One notable format which I relied on is the Tanzil project[1]. The project aims at providing a highly verified precise Qur'anic text. Qur'anic text is also available in the public domain in various format like MySQL database and XML. Multiple annotated corpora on the Qur'an are also available like the QAC and QurAna corpus I produced. In this chapter I have described a number of full or partial annotation projects on the Qur'an.

[1] Available online at http://tanzil.net

4.6.4 Standard Reference

The Qur'an has been a standard reference since its revelation. Apart from being the most vital religious reference to matters related to the daily life of Muslims, Arabic linguists has been referencing Qur'anic quotes for evidence of Arabic grammar topics. For example, I have done a phrase search on the classical book "Al of the most famous Arabian linguist and grammarian Sibawayh (died approx.. 796 CE) for the term "Allah said" and found 83 hits and all of them are citations from the Qur'an in support of certain grammar rule.

Regardless of the debate on the semantic meaning of the word 'corpus' and what merits to be called such, I think that employing corpus linguistic means to Qur'anic studies is beneficial to Qur'anic scholars, students and the general Muslim and non-Muslim community.

4.7 Summary

This chapter introduced a number of existing annotations of the Qur'anic texts. Some of the reviewed annotations have exhaustively covered the entire Qur'an – for example, our QurAna corpus or the corpus of pause marks of the Qur'an, and some are partial attempts, for example, syntactic annotation Treebank available at the Qur'anic Arabic Corpus website.

Haifa corpus and Qur'anic Arabic Corpus exhaustively analysed the morphology of every word of the Qur'an. However, QAC went through a rigorous verification process and adapted a collaborative platform for further evaluation.

QurAna picked up pronoun tags from QAC and annotated their referents. The creation of a concept ontology while tagging pronoun referents is a novel experiment I carried through this research. This approach proved helpful as the Qur'an is loaded with many pronouns.

This chapter included brief description of some partial annotation attempts to capture semantics of the Qur'an, like the Qur'anic semantic frames and Qur'anic prepositional verbs.

Finally, this chapter benchmarked the Qur'an against typical characteristics of a corpus: sampling and representativeness, finite size, machine-readable form and a standard reference.

Chapter 5
QurAna: Corpus of the Qur'an annotated with Pronominal Anaphora

5.1 Introduction

QurAna –or **Qur'an**ic **Ana**phora- is a corpus where all Qur'anic personal pronouns (over 24,000 pronouns) are tagged with information on their antecedents. The size of this corpus compares favourably with other available similar corpora. This chapter provides a detailed description of the annotation scheme, annotation process and usage of this corpus. For background linguistic information on pronominal anaphora in Arabic and in the Qur'an refer back to section 2.2. For a review of literature on anaphora resolution systems refer to section 3.2.

5.2 Annotation Scheme

5.2.1 Related annotation schema

The first anaphora annotation scheme is the Lancaster IBM project at UCREL (Garside et al. 1997). Under this scheme the antecedent (whether anaphor or cataphor) is enclosed in brackets and given an index number and the proform (i.e, the pronoun) is preceded by the 'REF' symbol with the index number along with either '<' or '>' symbol indicating the direction: either anaphora or cataphora. Figure 5.1 provides an example.

```
(6 the married couple 6) said that <REF=6 they were happy
with <REF=6 their lot.
```

Figure 5.1 – Example of annotation scheme from UCREL project (Garside et al. 1997)

This scheme was used to annotate part of the AP corpus consisting of around 100,000 words.

Another tagging schema is MUC-7 SGML schema (Hirschman and Chinchor 1997) which accompanied the task definition of the MUC-7 on coreference annotation. Figure 5.2 is a sample annotation from this corpus.

```
<COREF ID="100">Lawson Mardon Group Ltd.</COREF> said <COREF
ID="101" TYPE="IDENT" REF="100">it</COREF>…
```

Figure 5.2 – Example of annotation scheme from MUC-7 SGML schema (Hirschman and Chinchor 1997)

The GNOME project relies on an earlier general purpose annotation scheme called MATE. This scheme was designed keeping in mind a 'discourse model' and thus aimed at annotating 'discourse entities' and any co-reference to them. Under this scheme <de> is the main discourse element, and <link> is used to mark information about anaphoric relations using <anchor> elements. Figure 5.3 gives an example (Poesio 2004).

```
<de ID="de 01">we</de>'re gonna take <de ID="de 07"> the
engine E3 </de> and shove <de ID="de 08"> it </de> over to
<de ID="de 02">Corning</de>, hook <de ID="de 09"> it </de>
up to <de ID="de 03">the tanker car</de>...<link
href="coref.xml#id(de 07)" type="ident"> <anchor
href="coref.xml#id(de 08)"/> </link> <link
href="coref.xml#id(de 08)" type="ident"> <anchor
href="coref.xml#id(de_09)"/> </link>
```

Figure 5.3 – Example of annotation scheme from GNOME project (Poesio 2004)

AQA (Boldrini et al. 2009) is a multilingual anaphora annotation scheme that can be applied in machine learning for the improvement of Question Answering systems. This scheme has been used to annotate the CLEF 2008 corpus in Spanish. There are several markups used to specify anaphora type (e.g., pronominal, superficial, adverbial, ellipse and definite descriptions) and others to specify the relation type between anaphoric

expression and its direct or bridging antecedent. Figure 5.4 gives an example, where `<t>` = topic, `<subt>`=subtopic, `<q>`= question, `<de>`=discourse entity, `<link>`= anaphora, `<rel>`=relationship, `<status>`= sure or uncertain, `<ant>`=antecedent, `<refq>`=question-answer pair.

```
<t>
<q id="q538">
What was the name of the plane used by <de id="n52">John
Paul II</de> in <link rel="indir" status="ok" ant="q"
refq="q538" type="dd" ref="n52"> his travel</link> to the
USA in 1995?
</q>
<subt>
<q id="q539"> What instrument did Niccol Paganini play?
</q>
</subt>
</t>
```

Figure 5.4 - Example of annotation scheme from AQA project (Boldrini et al. 2009)

Using this scheme a pilot evaluation corpus was manually annotated out of the CLEF multilingual corpus with 600 questions: 200 for each English, Italian and Spanish with an average agreement of 87%.

(Hammami et al 2009) presents a corpus annotated with coreference chains for Arabic using a custom-designed XML-tool called AnATAr. This corpus is of size 77,457 words in size (very close to the size of the Qur'an) and includes newspaper articles, technical manual, a book on education and a novel. The scheme is adapted from (Tutin et al 2000) and is compatible with MUC scheme. Figure 5.5 below is an example output.

```
<s> … <exp id="e2" cat="Np" fc="sujet"> خديجة</exp></s>

<s> :حملت على عاتق<exp id="e3" cat="pln" dist="1" rec="true">

<ptr type="coref" src="e2" /> ها </exp>..
```

Figure 5.5 - Example of annotation scheme of Arabic anaphora corpus by (Hammami et al 2009)

5.2.2 QurAna Annotation Schema

QurAna preserves information about the location of both the pronoun and its antecedent. The annotation assumed the Qur'an segmentation available in the Qur'anic Arabic Corpus (QAC) and maintains the location information of pronouns through position of the word segments of the Qur'an. Consider for example the verse 67:5 below.

وَلَقَدْ زَيَّنَّا السَّمَاءَ الدُّنْيَا **بِمَصَابِيحَ** وَجَعَلْنَاهَا رُجُومًا لِّلشَّيَاطِينِ

And certainly / We have beautified / the heaven / nearest / with **lamps**, / and We have made them / (as) missiles / for the devils,

And verily We have beautified the world's heaven with lamps, and We have made them missiles for the devils,

Qur'anic Verse – 67 : 5

Figure 5.6 below is an excerpt of annotation of the above verse. Each word segment is denoted by its segment ID number using the XML tag: <seg id= xx> . These ID numbers corresponds to QAC segment numbers. The pronoun is indicated using the tag: <pron> which corresponds to QAC tag named 'PRON'. Moreover, each annotation further carries two attributes: a) the antecedent: ant-id and b) a concept id: concept-id. Note that when the antecedent is not available then always ant-id is assigned zero.

```
<seg id = 119860>ال</seg>
<seg id = 119861>دنيا</seg>
<seg id = 119862>ب</seg>
<seg id = 119863>مصابيح</seg>
<seg id = 119864>و</seg>
<seg id = 119865>جعل</seg>
<seg id = 119866><pron ant-id=0 concent-id=1>نا</pron></seg>
<seg id = 119867><pron ant-id = 119863 concept-id = 353>ها</pron></seg>
```

Figure 5.6 – Excerpt from verse 67:5 showing QurAna annotation schema

5.3 Annotation Process

Guided by the previous annotation schema, and by the nature of Arabic usage of pronouns, and in particular domain specific usage of pronouns in the Qur'an, I started the annotation process. First, the 128,000 word segments of the Qur'an were maintained in a MySQL database and unique IDs were assigned to each word segment. Next, QAC was used to identify those targeted segments that contain pronouns, and for each pronoun the starting and ending IDs of the text span that represents antecedents were recorded manually through forms developed using PHP scripting language. I chose QAC as it underwent various verification levels for accuracy of part-of-speech annotation, and is placed for online collaboration for further validation. I took the instances of pronouns that were tagged as 'PRON' in the QAC corpus. This tag covers all kind of personal pronouns: 1st, 2nd and 3rd persons; singular, dual and plural; connected pronouns (where the personal pronoun is suffixed with noun or verb) and

separate stand-alone pronouns. However, I left out demonstrative and relative pronouns as their number is less in the Qur'an (approx. 15%).

5.4 Concept Ontology from QurAna

As I progressed with tagging pronoun antecedents in the Qur'an, I maintained a concept ontology out of these antecedents. In the context of this thesis, pronoun referents are considered 'concept' entries of this ontology. Given this understanding, I have assigned a single concept for every pronoun in my scope. Even when the antecedent of a particular pronoun is absent in the Qur'an, still I have tagged this pronoun with the concept it refers to. Thus, a 'pronoun referent' and a 'concept' could be interchangeably used in this context. For example, QurAna has a concept entry called 'the messengers' referring to all instances where pronouns are used in the Qur'an to refer to 'messengers' in general. However, I also have separate concepts for individual messengers mentioned as pronouns in the Qur'an like 'Moses', 'Abraham', 'Jacob', 'Muhammad', etc.

In total I have gathered over 1050 concepts. For example, a concept in our list is "Qur'an" and all pronouns in the Qur'an referring to "Qur'an" were linked to this concept regardless of the presence or absence of the actual antecedent in the immediately surrounding text, or even various different names by which the concept "Qur'an" is referred to in the Qur'an: like (كتاب) *Kitab* (book), (ذكر) *Dhirk* (remembrance), (الفرقان) *al-Furqan* (the criteria), and such other attributes. Similarly, all pronouns referring to prophet Muhammad were under a single concept the actual pronoun antecedents might have different words in the Qur'an like (رسول) *Rasoul* (messenger) or (نبي) *Nabi'* (prophet). Keeping an ontology of concepts is very helpful for information retrieval and many other useful applications. Table 5.1 gives the 20 most frequently referred-to concepts in our list.

Freq.	Arabic	English
3061	الله	Allah
1145	الذين آمنوا	those who believe
1141	محمد	Prophet Muhammad
1110	الناس	Mankind
1073	الكافرين	(Kaafir) the infidels
912	المشركين	the polytheists
727	كفار قريش	the infidels of Quraish
655	المنافقين	the hypocrites
651	المؤمنين	the believers

549	بنى إسرائيل	Children of Israel
542	المسلمون	Muslims
360	موسى	Moses
288	اليهود	the Jews
221	آل فرعون	Pharaoh's folk
216	القرآن	the Qur'an
204	الإنسان	Mankind
202	أهل الكتاب	people of the Book
201	الأمم السابقة	the past nations
196	منكرو البعث	those who deny resurrection
190	إخوة يوسف	brothers of Joseph

Table 5.1 - 20 Most frequent concepts in the Qur'an

We call this collection of referents an ontology as the referents constitute the comprehensive set of nominal concepts found in the Qur'an. Other ontologies of the Qur'an exist, but are based on Qur'anic scholars' observations and intuitions about the core concepts in the Qur'an, rather than data-oriented extraction of nominal referents. For example the ontology used in the Qur'any search-by-concept tool (Abbas and Atwell 2012) is derived from the index terms in a scholarly analysis of the Qur'an.

As indicated earlier, access to books of Tafsir (scholarly comments) are important to resolve certain ambiguous cases, especially those instances where antecedents are absent. Consider for example verse 23:67 below.

مُسْتَكْبِرِينَ بِهِ سَامِرًا تَهْجُرُونَ

(Being) arrogant / about it, / conversing by night, / speaking evil." /

In scorn thereof. Nightly did ye rave together.

Qur'anic Verse - 23:67

The pronoun 'it' in this verse has no explicit antecedent mentioned and as such could refer to multiple entities like, Prophet Muhammad, the Qur'an or 'the house of Allah'. However, after consulting books of Tafsir, the majority votes for the last possibility. Similarly the pronoun 'him' in verse 22:15 below refers to Prophet Muhammad without any previous mention in the context.

مَن كَانَ يَظُنُّ أَن لَّن يَنصُرَهُ اللَّهُ فِي الدُّنْيَا وَالْآخِرَةِ فَلْيَمْدُدْ بِسَبَبٍ إِلَى السَّمَاءِ ثُمَّ لْيَقْطَعْ فَلْيَنظُرْ هَلْ يُذْهِبَنَّ كَيْدُهُ مَا يَغِيظُ

Whoever / [is] / thinks / that / not / Allah will help him / in / the world / and the Hereafter, / then let him extend / a rope / to / the sky, / then / let him cut off, / then let him see / whether / will remove / his plan / what / enrages. /

Whoso is wont to think (through envy) that Allah will not give him (Muhammad) victory in the world and the Hereafter (and is enraged at the thought of his victory), let him stretch a rope up to the roof (of his dwelling), and let him hang himself. Then let him see whether his strategy dispelleth that whereat he rageth!.

Qur'anic Verse - 22:15

The annotation of the pronoun 'him' in the above verse shows the vital role of *Tafsir* books in determining pronoun antecedents. Purely resorting to linguistic and syntactic means would tend to attach this pronoun with 'whoever' mentioned in the beginning of the verse, in which case the meaning would not seem appropriate to be attached with Allah.

This manual annotation was done by the thesis author and it took him over one year to annotate a total of 24,679 pronouns that cover the entire Qur'an.

5.5 Quality Assurance

QurAna is a product of semi-automated human annotation. As described earlier, scripts collected all instances of pronouns from QAC corpus and assigned the corresponding segment ID number automatically. Manual intervention from the human annotator involved for each identified pronoun:

a) find the text span of the antecedent and entering the segment IDs of the starting and ending phrases based on QAC segmentation

b) tag antecedent location as zero to indicate that the antecedent is either absent or its distance is far away

c) assign either an existing 'concept' or create a new concept as a referent for the pronoun at hand.

Following is a brief discussion on matters related to assuring the quality of annotation in terms of accuracy and ambiguity resolution for each of the above annotation decision.

QurAna annotation depended in attaining accuracy and resolving ambiguity on books of Tafsir accepted by the majority Sunni community of Muslims. Based on 2010 statistics, The Pew Forum on Religion and Public Life estimated the Sunni

population to be 1.4 billion among the 1.6 billion total Muslim population (88%)[1]. Among the books of Tafsir in the Sunni domain, QurAna annotator resorted to Tafsir Ibn Katheer (Ibn-Katheer 1372) (see section 6.2 for more information on this book).

The annotation is conducted by the author of this thesis. His Arabic proficiency compares with native speakers as he was enrolled in public Arabic schools in Saudi Arabia starting from elementary years. Moreover, the annotator has adequate exposure to the interpretation of the Qur'an and has memorized the entire Qur'an since early childhood. The annotator referred to the Tafsir book only in complicated and ambiguous cases where his native language instinct could not detect the referent.

The task of attaching a pronoun referent a 'concept' entry in the ontology is more judgemental by nature. However, to be consistent the following set of principles were followed during this task:

- If a referent is explicitly mentioned in the Qur'anic verse, ontology entry preserves the name of the referent.
- The above general rule is violated whenever the explicit referent in the Qur'an indicated a specific ontology concept. For example, often Qur'an refers to the concept of 'rain' as 'water send down from sky'. In such cases, QurAna registers these entries under 'rain' instead of 'water'.
- A concept in QurAna preserves one entry even if this concept refers in the Qur'an in multiple morphological or syntactic forms because of difference in number or gender or many morphological prefixes or suffixes.

5.6 Applications

Using PHP scripting and access to the annotated corpus captured as a MySQL database, a number of query pages were made online . Entering a verse number, a user can get all pronouns along with their antecedence and all concepts this verse has. Figure 5.7 below gives an example screenshot from the online query page. The actual verses are quoted in Arabic, however, the verse number leads to English – and potentially many other language translations through hyperlinks to an external site.

---

[1] "The Future of the Global Muslim Population", 2011, Pew Forum on Religion and Public Life. Available online at http://www.pewforum.org/The-Future-of-the-Global-Muslim-Population.aspx

gloss	Concept	Verse	Pronoun context	Antecedent	#
Allah	الله	38:29	أَمْ نَجْعَلُ ٱلْمُتَّقِينَ كَٱلْفُجَّارِ كِتَٰبٌ أَنزَلْنَٰهُ إِلَيْكَ مُبَٰرَكٌ لِّيَدَّبَّرُوٓا۟ ءَايَٰتِهِۦ وَلِيَتَذَكَّرَ أُو۟لُوا۟		1
the Qur'an	القرآن	38:29	أَمْ نَجْعَلُ ٱلْمُتَّقِينَ كَٱلْفُجَّارِ كِتَٰبٌ أَنزَلْنَٰهُ إِلَيْكَ مُبَٰرَكٌ لِّيَدَّبَّرُوٓا۟ ءَايَٰتِهِۦ وَلِيَتَذَكَّرَ أُو۟لُوا۟		2
Prophet Muhammad	محمد	38:29	نَجْعَلُ ٱلْمُتَّقِينَ كَٱلْفُجَّارِ كِتَٰبٌ أَنزَلْنَٰهُ إِلَيْكَ مُبَٰرَكٌ لِّيَدَّبَّرُوٓا۟ ءَايَٰتِهِۦ وَلِيَتَذَكَّرَ أُو۟لُوا۟ ٱلْأَلْبَٰبِ		3
mankind	الناس	38:29	كَٱلْفُجَّارِ كِتَٰبٌ أَنزَلْنَٰهُ إِلَيْكَ مُبَٰرَكٌ لِّ يَدَّبَّرُوٓا۟ ءَايَٰتِهِۦ وَلِيَتَذَكَّرَ أُو۟لُوا۟ ٱلْأَلْبَٰبِ وَوَهَبْنَا لِدَاوُۥدَ		4
the Qur'an	القرآن	38:29	كِتَٰبٌ أَنزَلْنَٰهُ إِلَيْكَ مُبَٰرَكٌ لِّيَدَّبَّرُوٓا۟ ءَايَٰتِهِۦ وَلِيَتَذَكَّرَ أُو۟لُوا۟ ٱلْأَلْبَٰبِ وَوَهَبْنَا لِدَاوُۥدَ سُلَيْمَٰنَ	كِتَٰبٌ	5

Figure 5.7 - Pronoun resolution of verse 38:29

Chapter 7.2 gives a more detailed account on the usage of this application and potential benefits as a text mining tool.

5.7 Quantitative Measures of QurAna

Following the process described above, the entire Qur'an was annotated. Table 5.2 below gives a quantitative account of key statistics of this corpus. Note that the 24,679 tagged pronouns include anaphoric as well as non-anaphoric cases, and relative and demonstrative pronouns are excluded.

Measure	Count	Description
# of word segments	127,795	A whitespace delimited word in Arabic could consist of multiple 'word segments' each equivalent to an English whitespace delimited word. In the Quran there are 77,430 white space delimited words. In order to compare with English, in this thesis I considered 'word segments' rather than just whitespace delimited words.
# of pronouns	24,679	This includes any PRON tagged by QAC, which excludes relative and demonstrative pronouns.
# of third person pronouns	11,544	Representing around 47% of Qur'anic pronouns.
# of Qur'anic Verses	6,236	A Qur'anic verse may vary greatly in size. While a verse like 2:282 contains 213 word segments, others like verses 2:1 and 36:1 contain just one word segment.
Average Distance between pronoun and antecedent	30 word segments	This measurement was enabled as QurAna keeps track of ID number of both a pronoun and the word segment of the antecedent.
% of antecedents within the same verse as the pronoun	56%	This measurement was enabled as QurAna keeps track of ID number of both a pronoun and the word segment

		of the antecedent. Tracing these ID numbers, QAC allows checking against the verse number of both a pronoun and its antecedence.
% availability of antecedents	54%	During manual annotation, QurAna annotator roughly checks availability of antecedence within the same page (approx.. 200 words window), except in case of long stories (like addressing 'the Children of Israel' in chapter 2. It is noted that a well-known category of non-availability of antecedence in the Qur'an is addressing the Prophet Muhammad as 2nd person pronoun (over 1,000 instances).
Total number of concepts	1054	Here 'concepts' refers to pronoun referents. When antecedents are missing, still pronouns are tagged against a concept.

Table 5.2 – Quantitative measures from QurAna corpus

Among the total pronouns tagged in the QurAna corpus only 90 instances (0.3%) showed cataphor relation where the antecedents were mentioned after the pronoun. Although in certain cases the antecedent is mentioned way back, in the majority of cases they are found within 200 word segments from the pronoun. Among the 13,158 pronouns which have antecedents, only 2,309 (17.5%) antecedents matched with the nearest preceding noun. Considering the whole population of pronouns only 9% of antecedents are captured correctly when attached with the nearest preceding noun. Among the total 2nd person singular pronouns 27% of them referred to Prophet Muhammad.

This corpus is made public for both online query – as described in the previous section 5,4 - or for download at the following webpage: http://www.textminingtheQuran.com/wiki/Pronoun_Reference_in_the_Quran

5.8 Challenges and future improvement

We have encountered a number of challenges while pursuing this task. Often, the distance between the pronoun and its antecedent is very long. This is evident more in the case of long stories, where the main characters might be mentioned only once at the beginning and all subsequent references are done through pronouns. For example, in Chapter 2 of the Qur'an, a series of verses addressed the 'Children of Israel' where explicit mention is made at the beginning of the dialogue but most subsequent references are made through 2nd person

pronouns sometimes as far as 33 verses away. Also, as our annotation scheme does not allow discontinuous antecedents or multiple antecedents, in such cases I had to include as antecedents the whole text span, resulting in some compound concepts.

Often the Qur'an makes grammatical shifts deliberately for various purposes (for example to draw attention), and as a result the number or person agreement between the pronoun and the antecedent is violated. Consider for example verse 65:1 where the singular noun antecedent 'prophet' disagrees with the plural 2nd person pronoun used (you):

يَا أَيُّهَا النَّبِيُّ إِذَا طَلَّقْتُمُ النِّسَاءَ فَطَلِّقُوهُنَّ لِعِدَّتِهِنَّ وَأَحْصُوا الْعِدَّةَ
O Prophet! / When / you divorce / [the] women, / then divorce them / for their waiting period,
O Prophet! When ye (men) put away women, put them away for their (legal) period and reckon the period,

Qur'anic Verse – 65:1

There were a number of challenges faced while tagging pronouns with a concept name. Often a decision to create a new specific concept or maintain an already available generic concept was required. For example, in verse 67:5 the word 'lamp' was used to mean 'stars', and hence pronouns could be tagged with either of these two concepts. In this particular case, I decided to tie the pronoun to the concept 'star' rather than the concept "lamp" so that all verses referring to 'star' can be linked plus as I keep reference to actual antecedence, I still can retrieve that stars are referred to in the Qur'an as lamps.

وَلَقَدْ زَيَّنَّا السَّمَاءَ الدُّنْيَا بِمَصَابِيحَ وَجَعَلْنَاهَا رُجُومًا لِّلشَّيَاطِينِ
And certainly / We have beautified / the heaven / nearest / with lamps, / and We have made them / (as) missiles / for the devils,
And verily We have beautified the world's heaven with lamps, and We have made them missiles for the devils,

Qur'anic Verse – 67:5

QurAna is characterized by: (a) comparatively large number of pronouns tagged with antecedent information, and (b) maintenance of an ontological concept list out of these antecedents. I have shown useful applications of this corpus. This corpus is the first of its kind considering Classical Arabic text, and would find interesting applications for Modern Standard Arabic as well, as is detailed in section 1.3.

5.9 Summary

This chapter discussed in detail QurAna corpus which captures annotation of over 24,000 Qur'anic pronouns with their antecedents and maintains in parallel an ontology of Qur'anic concepts from these antecedents. The annotation schema employed for building QurAna is comparable to other schema designed for similar tasks like the UCREL schema or MUC-7 SGML schema.

The chapter described how QAC was integrated with the annotation process and how available scholarly comments on the Qur'an was helpful in resolving ambiguous cases. This corpus along with concept ontology was incorporated into online application where users can query this corpus.

Finally, this chapter discussed some challenges faced during annotation process and future improvements that could enhance QurAna.

Chapter 6
QurSim: A corpus for evaluation of relatedness in short texts

6.1 Introduction

The ability to quantify computationally semantic relatedness of natural language short texts has many interesting applications such as: words sense disambiguation, information extraction and retrieval, automatic indexing, lexical selection, text summarization, automatic correction of word errors, and word and text clustering. Although this task is complex computationally, humans routinely perform semantic relatedness tasks readily both at word level, e.g. between the words “cat” and “mouse”, or at phrase and text level, e.g. between “drafting a letter” and “writing an email message”. This task has been fairly natural for humans because they can associate a huge amount of background experience and external domain concepts, whereas computational methods lack this smart association mechanism with related external sources.

In this chapter I describe QurSim, a corpus of short texts marked with relatedness information judged by human domain experts. The degree of relatedness between texts in this corpus varies greatly: although there are instances where lexical matching is evident between the terms in a pair of related texts, the majority of instances require deep semantic analysis and domain specific world knowledge in order to relate the two texts in the pair.

Our objective in collecting this dataset is to provide evaluation and training – and perhaps a gold-standard resource for researchers in the field of computational semantic similarity and relatedness analysis in natural language texts. Following is a detailed description of the dataset, the scholarly work from which the dataset was compiled, the compilation process, some applications of this dataset, evaluation, challenges and some future enhancements. For an introduction on text similarity and relatedness in the Qur’an refer to section 2.3. For a literature review on computational analysis of similarity and relatedness between words and short texts refer to section 3.3.

6.2 Book of Tafsir by Ibn Katheer

Ismail Ibn Katheer was a Muslim scholar who died in 1373 CE, well known for his classic book of Qur'an commentary (or Tafsir in Arabic). This book is one of the most widely cited commentaries of the Qur'an. Ibn Katheer followed a regular methodology when commenting on a verse, which he made clear at the introduction of his book (see section 2.5.2 for more details). Firstly, he discusses other related verses explaining the current verse. Often, when a certain verse covers a subject briefly, there might be many other verses that cover other aspects of this subject; see section 2.3.1 for a number of examples. Secondly, he refers to traditions and sayings of the Prophet Muhammad (i.e., Hadith). Thirdly, he cites opinions of Sahabah (i.e., companions of the Prophet) on this verse, especially those who are well known for their knowledge of the Qur'an like Ibn Abbas and Ibn Masoud.

We exploited this methodology for the purpose of creating a dataset of related verses. I understand that Ibn Katheer never claimed to exhaustively cite all related verses when commenting on a particular verse, nor did he always observe the commutative property of relatedness, i.e., if verse y was cited while commenting on verse x, then x should also appear at the commentary page of verse y. These observations allowed us to expand the original list of related verses beyond what is found in Ibn Katheer.

6.3 Compilation process

Tafsir Ibn Katheer is available online at several websites. I chose the online version available at the official website of the King Fahd Complex for the Printing of the Holy Qur'an . After observing the structured format used in this site for displaying this Tafsir, I developed scripts to extract verse pairs automatically. My script automatically retrieved the URL of a given verse and extracted the chapter and verse numbers of all other verses mentioned in the context of a given verse.

After the initial compilation of the dataset, manual intervention was necessary to clean up some inconsistencies, as well as to adjust correct pairing of verses, because in the original Ibn Katheer's Tafsir, a group of verses are discussed at a time, while our dataset contains only verse pairs. Through this process I collected a total of 7,679 pairs of single verses which were then fed into relational database tables using MySQL.

6.3.1 Dataset filtration and extension

Upon initial investigation of the dataset, the annotator realized –based on native language instinct as well as knowledge of the computational similarity task- that a second manual check was required to filter the dataset further for it to be useful for the intended computational analysis tasks. While commenting on a particular verse, Ibn Katheer's discussion might lead to a distant topic, making the task of computation of the relatedness almost impossible. The annotator made decision to brand such cases as 'weakly related' verse pairs.

Consider for example, the following pair from our dataset, where no obvious relation is found before reading the context in Ibn Katheer:

إِنَّ اللَّهَ لَا يَسْتَحْيِي أَن يَضْرِبَ مَثَلًا مَّا بَعُوضَةً فَمَا فَوْقَهَا ۚ فَأَمَّا الَّذِينَ آمَنُوا فَيَعْلَمُونَ أَنَّهُ الْحَقُّ مِن رَّبِّهِمْ ۖ وَأَمَّا الَّذِينَ كَفَرُوا فَيَقُولُونَ مَاذَا أَرَادَ اللَّهُ بِهَٰذَا مَثَلًا
Indeed, / Allah / (is) not / ashamed / to / set forth / an example / (like) even / (of) a mosquito / and (even) something / above it. / Then as for / those who / believed, / [thus] they will know / that it / (is) the truth / from / their Lord. / And as for / those who / disbelieved / [thus] they will say / what / (did) intend / Allah / by this / example?
Lo! Allah disdaineth not to coin the similitude even of a gnat. Those who believe know that it is the truth from their Lord; but those who disbelieve say: What doth Allah wish (to teach) by such a similitude?

Qur'anic Verse – 2:26

فَلَمَّا نَسُوا مَا ذُكِّرُوا بِهِ فَتَحْنَا عَلَيْهِمْ أَبْوَابَ كُلِّ شَيْءٍ حَتَّىٰ إِذَا فَرِحُوا بِمَا أُوتُوا أَخَذْنَاهُم بَغْتَةً فَإِذَا هُم مُّبْلِسُونَ
So when / they forgot / what / they were reminded / of [it], / We opened / on them / gates / (of) every / thing, / until / when / they rejoiced / in what / they were given, / We seized them / suddenly / and then / they / (were) dumbfounded. /
Then, when they forgot that whereof they had been reminded, We opened unto them the gates of all things till, even as they were rejoicing in that which they were given, We seized them unawares, and lo! they were dumbfounded.

Qur'anic Verse – 6:44

Ibn Katheer drew an analogy between the situation of a gnat (or mosquito) who overfeeds itself till death (the first verse in the pair), and those people who wrongly over-enjoy the provision of this world till God's punishment befalls them (the second verse in the pair).

Considering these kinds of examples, the entire dataset was manually checked and - instead of completely removing these pairs - a special 'not obvious' flag was placed against 883 such cases. With the remaining 6,796 semantically related verse pairs, we believed further distinctions in the degree of relatedness were needed if the dataset were to be used for training learning algorithms. Consider for example verse 78:20 below.

وَسُيِّرَتِ الْجِبَالُ فَكَانَتْ سَرَابًا

And are moved / the mountains / and become / a mirage. /

And the hills are set in motion and become as a mirage.

Qur'anic Verse – 78:20

Ibn Katheer cited the following three consecutive verses in his commentary on 78:20. While the first verse 20:105 is strongly related to 78:20, the other two complete the picture in the context.

وَيَسْأَلُونَكَ عَنِ الْجِبَالِ فَقُلْ يَنسِفُهَا رَبِّي نَسْفًا

And they ask you / about / the mountains, / so say, / "Will blast them / my Lord / (into) particles. /

They will ask thee of the mountains (on that day). Say: My Lord will break them into scattered dust. [20:105]

فَيَذَرُهَا قَاعًا صَفْصَفًا

Then He will leave it, / a level / plain. /

And leave it as an empty plain, [20:106]

لَّا تَرَىٰ فِيهَا عِوَجًا وَلَا أَمْتًا

Not / you will see / in it / any crookedness / and not / any curve." /

Wherein thou seest neither curve nor ruggedness. [20:107]

Qur'anic Verses – 20:105-107

Thus, a second scrutiny of the dataset resulted in assigning two levels of degree of relatedness: level 2 (in total 3,079 pairs) represents strong relations as between verses 78:20 and 20:105, and level 1 (total 3,718 pairs) represents weaker relations as between 78:20 and 20:106 above. Manual filtration of all levels described above was performed by the author.

We suggest two ways in which this dataset could be extended: (a) for a pair of strongly related verses <x,y> (i.e., level 2) the pair <y,x> should be included if not already in the dataset. (b) Consider a related pair <x,y>, if <y,z> is also strongly related, then both <x,z> and <z,x> could be added as well. However, since a verse – especially if large in size- can mention several different things, generalizing this transitivity property would result in a large number of unrelated associations. The QurAna dataset does not contain these suggested extended pairs, as they could be computed automatically.

وَسُيِّرَتِ الْجِبَالُ فَكَانَتْ سَرَابًا

And are moved / the mountains / and become / a mirage. /

And the hills are set in motion and become as a mirage.

Qur'anic Verse - 78:20

وَيَوْمَ نُسَيِّرُ الْجِبَالَ

And the Day / We will cause (to) move / the mountains /

And (bethink you of) the Day when we remove the hills..

Qur'anic Verse - 18:47

وَتَسِيرُ الْجِبَالُ سَيْرًا

And will move away, / the mountains / (with an awful) movement /

And the mountains move away with (awful) movement

Qur'anic Verse - 52:10

As an illustration, consider the three verses above. I find that <78:20, 18:47> is a level 2 pair in our dataset. However, <18:47, 78:20> is not found but could be added as a new pair. Similarly, I notice that <18:47, 52:10> is a level 2 pair in the

dataset, however, the pair <78:20, 52:10> was not considered by Ibn Katheer neither was the pair <52:10,78:20>, and both could be added as strongly related verses.

The Qur'anic Arabic Corpus (QAC) gives the root of each word, so I used this to count the availability of matching lexical roots in all paired verses in our dataset.

The dataset was made available for download as XML file containing the format shown in figure 6.1 below.

```
<column name="uid">1</column>
<column name="ss">1</column>
<column name="sv">1</column>
<column name="ts">1</column>
<column name="tv">2</column>
<column name="common">0</column>
<column name="relevance">2</column>
```

Figure 6.1 – A sample QurSim XML representation

The XML representation above shows that the related pair of verses `<ss:sv, ts:tv>` has a `common` number of matching lexical roots and are related with a degree of `relevance`, which could be 0,1 or 2. The available file contains the original 7,679 pairs, while the extension of the dataset could be made computationally following the logic described above. I kept only reference to chapter and verse numbers since most electronic version of the original Qur'an text as well as its translations maintain these references.

6.4 Quality Assurance

QurSim relied on Tafsir Ibn Katheer to compile the dataset of related verses. As discussed earlier (section 5.5), this source is considered a classic source of

Qur'an interpretation by Sunni Muslims who constitute over 88% of Muslim population.

Pairs of related verses in the QurSim dataset was compiled by automatic scripts that traversed online version of Tafsir Ibn Katheer to produce these pairs. Manual intervention by the human annotator involved the following decisions:

a) Verifying the correct pairing of a verse with its related verse.

b) flagging certain verse pairs as weakly related, using annotator's native speaker's knowledge of language and the world. See detailed discussion in section 6.3.1

c) flagging certain verse pairs as strongly related, using annotator's native speaker's knowledge of language, world and NLP capabilities. This strongly related category was judged by annotator's familiarity of the capabilities of automatic computational methods to detect similarity in short texts. All verse pairs not flagged in (b) or (c) above are considered the default relation of related verses as considered so by Ibn Katheer.

The annotation is conducted by the author of this thesis. His Arabic proficiency compares with native speakers as he was enrolled in public Arabic schools in Saudi Arabia starting from elementary years. Moreover, the annotator has adequate exposure to the interpretation of the Qur'an and has memorized the entire Qur'an since early childhood.

6.5 Applications

The QurSim dataset has been captured as MySQL in order to enable web queries and visualization. I have created online query pages where a user inputs a verse number and is returned with both directly and indirectly related verses, in Arabic and English, along with the degree of relatedness and common roots as shown in figure 6.2. Moreover, thanks to integration with QurAna (see Chapter 5), we provided information on concepts as antecedents of pronouns in each verse, as well as a concept cloud (see figure 6.3) from all verses, given at the end to give the user an idea of the major concepts involved. Chapter 7.3 gives a detailed account of the features of this application and the role it can play as a text mining tool.

Following are 5 verses directly related to 7:187 from Ibn Kathir:

No.	Arabic	English	Common Roots	Level
33:63	يَسْأَلُكَ النَّاسُ عَنِ السَّاعَةِ قُلْ إِنَّما عِلْمُها عِندَ اللَّهِ وَما يُدريكَ لَعَلَّ السَّاعَةَ تَكونُ قَريبًا	Men ask thee of the Hour. Say: The knowledge of it is with Allah only. What can convey (the knowledge) unto thee? It may be that the Hour is nigh. **Pronoun Referents:** Prophet Muhammad, the Hour,	11	2
31:34	إِنَّ اللَّهَ عِندَهُ عِلمُ السَّاعَةِ وَيُنَزِّلُ الغَيثَ وَيَعلَمُ ما فِي الأَرحامِ وَما تَدري نَفسٌ ماذا تَكسِبُ غَدًا وَما تَدري نَفسٌ بِأَيِّ أَرضٍ تَموتُ إِنَّ اللَّهَ عَليمٌ خَبيرٌ	Lo! Allah! With Him is knowledge of the Hour. He sendeth down the rain, and knoweth that which is in the wombs. No soul knoweth what it will earn to-morrow, and no soul knoweth in what land it will die. Lo! Allah is Knower, Aware. **Pronoun Referents:** Allah,	7	2
42:18	يَستَعجِلُ بِها الَّذينَ لا يُؤمِنونَ بِها وَالَّذينَ ءامَنوا مُشفِقونَ مِنها وَيَعلَمونَ أَنَّهَا الحَقُّ أَلا إِنَّ الَّذينَ يُمارونَ فِي السّاعَةِ لَفي ضَلٰلٍ بَعيدٍ	Those who believe not therein seek to hasten it, while those who believe are fearful of it and know that it is the Truth. Are not they who dispute, in doubt concerning the Hour, far astray? **Pronoun Referents:** the Hour, (Kaafir) the infidels, the believers , those who deny resurrection,	4	2
79:42	يَسْأَلونَكَ عَنِ السّاعَةِ أَيّانَ مُرسىٰها	They ask thee of the Hour: when will it come to port? **Pronoun Referents:** the Hour, Prophet Muhammad, the infidels of Quraish,	4	2
21:38	وَيَقولونَ مَتىٰ هٰذَا الوَعدُ إِن كُنتُم صٰدِقينَ	And they say: When will this promise (be fulfilled), if ye are truthful? **Pronoun Referents:** the infidels of Quraish, Prophet Muhammad and the believers,	2	2

Figure 6.2 - Verses directly related to 7:187

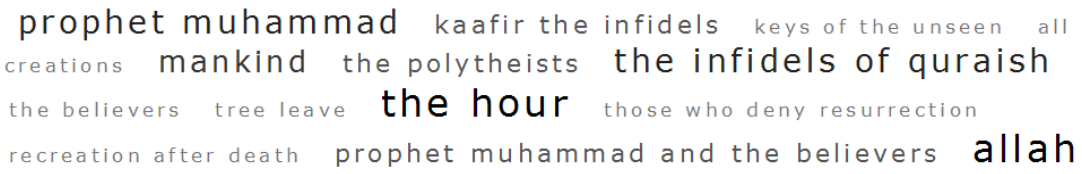


Figure 6.3 - Concept cloud from pronoun referents of all related verses to 7:187

6.6 Challenges

As Qur'anic verses vary in size I run into two different problems: 1) Those verses that are long may cover several topics and hence, pairing the whole verse with another verse reflects only a partial relation 2) those verses that are very small share with adjacent verses a single topic, and again in this case the one-to-one pairing with another verse is not appropriate. As an example for the first point, consider verses related to verse 2:255 below. This is a relatively long verse containing 10 short sentences covering different aspect of Allah's attributes and quality. Ibn Katheer links this verse with 15 different verses.

اللَّهُ لَا إِلَٰهَ إِلَّا هُوَ الْحَيُّ الْقَيُّومُ ۚ لَا تَأْخُذُهُ سِنَةٌ وَلَا نَوْمٌ ۚ لَّهُ مَا فِي السَّمَاوَاتِ وَمَا فِي الْأَرْضِ ۗ مَن ذَا الَّذِي يَشْفَعُ عِندَهُ إِلَّا بِإِذْنِهِ ۚ يَعْلَمُ مَا بَيْنَ أَيْدِيهِمْ وَمَا خَلْفَهُمْ ۖ وَلَا يُحِيطُونَ بِشَيْءٍ مِّنْ عِلْمِهِ إِلَّا بِمَا شَاءَ ۚ وَسِعَ كُرْسِيُّهُ السَّمَاوَاتِ وَالْأَرْضَ ۖ وَلَا يَئُودُهُ حِفْظُهُمَا ۚ وَهُوَ الْعَلِيُّ الْعَظِيمُ

Allah - / (there is) no / God / except / Him, / the Ever-Living, / the Sustainer of all that exists. / Not / overtakes Him / slumber / [and] not / sleep. / To Him (belongs) / what(ever) / (is) in / the heavens / and what(ever) / (is) in / the earth. / Who / (is) the one / who / can intercede / with Him / except / by / He knows / what / (is) / before them / and what / (is) behind them. / And not / they encompass / anything / of / His Knowledge / except / [of] what / He willed. / Extends / His Throne / (to) the heavens / and the earth. / And not / tires Him / (the) guarding of both of them. / And He / (is) the Most High, / the Most Great. /

Allah! There is no deity save Him, the Alive, the Eternal. Neither slumber nor sleep overtaketh Him. Unto Him belongeth whatsoever is in the heavens and whatsoever is in the earth. Who is he that intercedeth with Him save by His leave? He knoweth that which is in front of them and that which is behind them, while they encompass nothing of His knowledge save what He will. His throne includeth the heavens and the earth, and He is never weary of preserving them. He is the Sublime, the Tremendous.

Qur'anic Verse – 2:255

As an example on the second point, consider verse 11:97 below. Ibn Katheer referred to six consecutive small verses as related to this verse. Since I paired one single verse with another, in our dataset this relation is represented by six pairs <11:97,79:21>, <11:97, 79:22>, … <11:97,79:26>. Note how the last pair <11:97, 79:26> is very weakly related when taken in isolation.

إِلَىٰ فِرْعَوْنَ وَمَلَئِهِ فَاتَّبَعُوا أَمْرَ فِرْعَوْنَ ۖ وَمَا أَمْرُ فِرْعَوْنَ بِرَشِيدٍ

To / Firaun / and his chiefs, / but they followed / (the) command of Firaun, / and not / (the) command of Firaun / / was right. /

Unto Pharaoh and his chiefs, but they did follow the command of Pharaoh, and the command of Pharaoh was no right guide.

Qur'anic Verse – 11:97

فَكَذَّبَ وَعَصَىٰ ثُمَّ أَدْبَرَ يَسْعَىٰ فَحَشَرَ فَنَادَىٰ فَقَالَ أَنَا رَبُّكُمُ الْأَعْلَىٰ فَأَخَذَهُ اللَّهُ نَكَالَ الْآخِرَةِ وَالْأُولَىٰ إِنَّ فِي ذَٰلِكَ لَعِبْرَةً لِّمَن يَخْشَىٰ

But he denied / and disobeyed. / Then / he turned his back, / striving, / And he gathered / and called out, / Then he said, / "I am / your Lord, / the Most High." / So seized him / Allah / (with) an exemplary punishment / (for) the last / and the first. / Indeed, / in / that / surely (is) a lesson / for whoever / fears. /

[21]But he denied and disobeyed, [22]Then turned he away in haste, [23]Then gathered he and summoned, [24]And proclaimed: " I (Pharaoh) am your Lord the Highest." [25] So Allah seized him (and made him) an example for the after (life) and for the former. [26] Lo! herein is indeed a lesson for him who feareth.

Qur'anic Verses – 79:21-26

Another challenge I faced is when Ibn Katheer elaborates on a particular word from a verse and brings in different verses in the course of explanation. These cited verses might not seem related without relating back to the context made in Ibn Katheer. For example consider the verse 11:8 where the word "Ummah" was mentioned, which means a "nation". However, in the Qur'an this word can have other less frequently used meanings like "a leader" or "a short period of time". Here Ibn Katheer cites references to all other verses in the Qur'an where this word is used to mean things other than a "nation".

6.7 Future Improvements

The dataset could be improved further. As Qur'anic verses vary in size, a pair of two large size verses might relate based on a smaller phrase within these verses. Such instance of pairs could be cropped so only related phrases are preserved. Books of Tafsir other than Ibn Katheer could be consulted to increase the size of our dataset. Traditions of Prophet Muhammad narrated to explain verses could also be incorporated from Ibn Katheer to enrich this dataset.

Computational analysis of text relatedness is a growing research area. The lack of proper evaluation datasets stands as a major obstacle for progress in this field. Because of the availability of machine readable Qur'an translations in multiple languages, QurSim can potentially contribute in producing quality datasets in multiple languages and with minimum effort.

6.8 Summary

This chapter presented QurSim: a large corpus created from the original Qur'anic text, where semantically similar or related verses are linked together. This dataset can be used for evaluation of paraphrase analysis and machine translation tasks. QurSim is characterised by: (1) superior quality of relatedness assignment; as QurSim has incorporated relations marked by well-known domain experts, this dataset could thus be considered a gold standard corpus for various evaluation tasks, (2) the size of QurSim; over 7,600 pairs of related verses are collected from scholarly sources with several levels of degree of relatedness. This dataset could be extended to over 13,500 pairs of related verses observing the commutative property of strongly related pairs.

This dataset was incorporated into online query pages where users can visualize for a given verse a network of all directly and indirectly related verses. Empirical experiments showed that only 33% of related pairs shared root words, emphasising the need to go beyond common lexical matching methods, and incorporate -in addition- semantic, domain knowledge, and other corpus-based approaches.

This chapter concluded with describing some challenges faced during the compilation process and suggested some ways to improve QurSim in future.

Chapter 7
Text Mining Application on the Qur'an

7.1 Qur'anic Concordancer: QurConcord

A concordance is considered the single most important tool available to the corpus linguist (McEnery and Hardie 2012). It enables retrieving evidence from a corpus displayed in one-example-per-line format with the context before and after each example.

A concordancer for the Arabic Qur'an was designed, implemented and made available on-line. Figure 7.1 shows a screenshot of a sample concordance lines for the input word "كتاب"/*kitab* or "book".

2:2	لَا رَيْبَ فِيهِ هُدًى لِّلْمُتَّقِينَ ٱلَّذِينَ يُؤْمِنُونَ	ٱلْكِتَٰبُ	غَيْرِ ٱلْمَغْضُوبِ عَلَيْهِمْ وَلَا ٱلضَّآلِّينَ ال۪مٓ ذَٰلِكَ	1
2:44	أَتَأْمُرُونَ ٱلنَّاسَ بِٱلْبِرِّ وَتَنسَوْنَ أَنفُسَكُمْ وَأَنتُمْ تَتْلُونَ	ٱلْكِتَٰبَ	أَفَلَا تَعْقِلُونَ وَٱسْتَعِينُوا۟ بِٱلصَّبْرِ وَٱلصَّلَوٰةِ وَإِنَّهَا لَكَبِيرَةٌ	2
2:53	بَعْدِ ذَٰلِكَ لَعَلَّكُمْ تَشْكُرُونَ وَإِذْ ءَاتَيْنَا مُوسَى	ٱلْكِتَٰبَ	وَٱلْفُرْقَانَ لَعَلَّكُمْ تَهْتَدُونَ وَإِذْ قَالَ مُوسَىٰ لِقَوْمِهِۦ	3
2:78	يُسِرُّونَ وَمَا يُعْلِنُونَ وَمِنْهُمْ أُمِّيُّونَ لَا يَعْلَمُونَ	ٱلْكِتَٰبَ	إِلَّآ أَمَانِىَّ وَإِنْ هُمْ إِلَّا يَظُنُّونَ فَوَيْلٌ	4
2:79	وَإِنْ هُمْ إِلَّا يَظُنُّونَ فَوَيْلٌ لِّلَّذِينَ يَكْتُبُونَ	ٱلْكِتَٰبَ	بِأَيْدِيهِمْ ثُمَّ يَقُولُونَ هَٰذَا مِنْ عِندِ ٱللَّهِ	5
2:85	تُفَٰدُوهُمْ وَهُوَ مُحَرَّمٌ عَلَيْكُمْ إِخْرَاجُهُمْ أَفَتُؤْمِنُونَ بِبَعْضِ	ٱلْكِتَٰبِ	وَتَكْفُرُونَ بِبَعْضٍ فَمَا جَزَآءُ مَن يَفْعَلُ ذَٰلِكَ	6
2:87	ٱلْعَذَابُ وَلَا هُمْ يُنصَرُونَ وَلَقَدْ ءَاتَيْنَا مُوسَى	ٱلْكِتَٰبَ	وَقَفَّيْنَا مِنۢ بَعْدِهِۦ بِٱلرُّسُلِ وَءَاتَيْنَا عِيسَى ٱبْنَ	7
2:89	ٱللَّهُ بِكُفْرِهِمْ فَقَلِيلًا مَّا يُؤْمِنُونَ وَلَمَّا جَآءَهُمْ	كِتَٰبٌ	مِّنْ عِندِ ٱللَّهِ مُصَدِّقٌ لِّمَا مَعَهُمْ وَكَانُوا۟	8
2:101	لِّمَا مَعَهُمْ نَبَذَ فَرِيقٌ مِّنَ ٱلَّذِينَ أُوتُوا۟	ٱلْكِتَٰبَ	كِتَٰبَ ٱللَّهِ وَرَآءَ ظُهُورِهِمْ كَأَنَّهُمْ لَا يَعْلَمُونَ	9
2:101	مَعَهُمْ نَبَذَ فَرِيقٌ مِّنَ ٱلَّذِينَ أُوتُوا۟ ٱلْكِتَٰبَ	كِتَٰبَ	ٱللَّهِ وَرَآءَ ظُهُورِهِمْ كَأَنَّهُمْ لَا يَعْلَمُونَ وَٱتَّبَعُوا۟	10

Figure 7.1 – First 10 concordance lines for the input word "كتاب"

First, the user inputs a surface word in un-vocalized Arabic plain text. Next, a number of background processing is done where: a) a search is made on all instances of this word in an un-vowelized version of the Qur'an, and b) for each found word the corresponding lemma or root word –whichever is present- is taken from Qur'anic Arabic Corpus (QAC) as the target examples, and c) finally, the instances are rendered through the website.

Following are some features of this custom made application.

- QurConcord is rendered online[1] using PHP scripting language MySQL backend database.
- An online wiki page is maintained documenting the usage of QurConcord[2].
- The online page is designed using PHP scripting language and query word is embedded in the URL link, enabling getting data through URL.
- QurConcord brings 8 words before and after the query word
- The results are sorted based on increasing order of Qur'anic chapter numbers.
- Qur'anic reference is associated with each example as a link that leads to the corresponding verse page at Qur'anic Arabic Corpus (QAC).
- Each focus word is depicted in bold, and on pointing over by mouse a detailed morphological description is shown as a tooltip box, as shown in figure 7.1 above.
- Each word in an example line can be pointed at by mouse to reveal its POS taken from QAC.
- Each work in an example line can be clicked to reveal the concordance line of this word.

To the best of our knowledge, no custom made concordancer for the Qur'an exists today. QurConcord can be a very useful tool in the hand of Qur'anic linguistic researchers. Rendering this concordance online, I anticipate novel application of this tool by Qur'anic researchers. I already received a number of complimentary e-mails since launching this tool. One useful application could be in building specialized Qur'anic semantic frames (Fillmore 1976). A concordancer helps in discovering the various frame elements. Consider for example figure 7.2 below, where some representative concordance lines for this verb from the Qur'an (Muhammad and Atwell 2009) are rendered after translation. I notice that, apart from the 'ingestion of food' sense of the word 'eat', Qur'an also refers to figurative sense of 'eating money unlawfully' as in example B.

---

[1] Available at http://textminingtheQuran.com/php/con.htm

[2] Available at http://www.textminingtheQuran.com/wiki/QurConcord

A	the sea to be of service that ye	**eat**	fresh meat from thence	16:14
B	And	**eat**	not up your property among	2:188
C	Would one of you love to	**eat**	the flesh of his dead brother?	49:12
D	seven fat kine which seven lean were	**eating**		12:43
E	seven hard years which will	**eat**	all that ye have prepared for them	12:48
F	they	**eat**	into their bellies nothing else than fire	2:177
G		**eat**	of unlawful	5:42

Figure 7.2 – sample concordance lines for the verb ‘eat’ in the Qur’an (Muhammad and Atwell 2009)

In future, QurConcord can be greatly improved by incorporating more search and display features such as those detailed in (Hardie forthcoming) including support for regular expression, multi-word and suffix search.

7.2 QurAna: Qur’anic Pronoun Referents Application

This online application was built as a query and visualization tool for the QurAna corpus discussed in detail at chapter 5.

This application uses PHP scripting and access to the annotated corpus captured as a MySQL database, a number of query pages were made online[1]. Entering a verse number, a user can get all pronouns along with their antecedence and all concepts this verse has. Figure 7.3 below gives an example screenshot from the online query page. The actual verses are quoted in Arabic, however, the verse

[1] Available online at http://www.textminingtheQuran.com/apps/pron.php

number leads to English – and potentially many other language translations through hyperlinks to an external site[1].

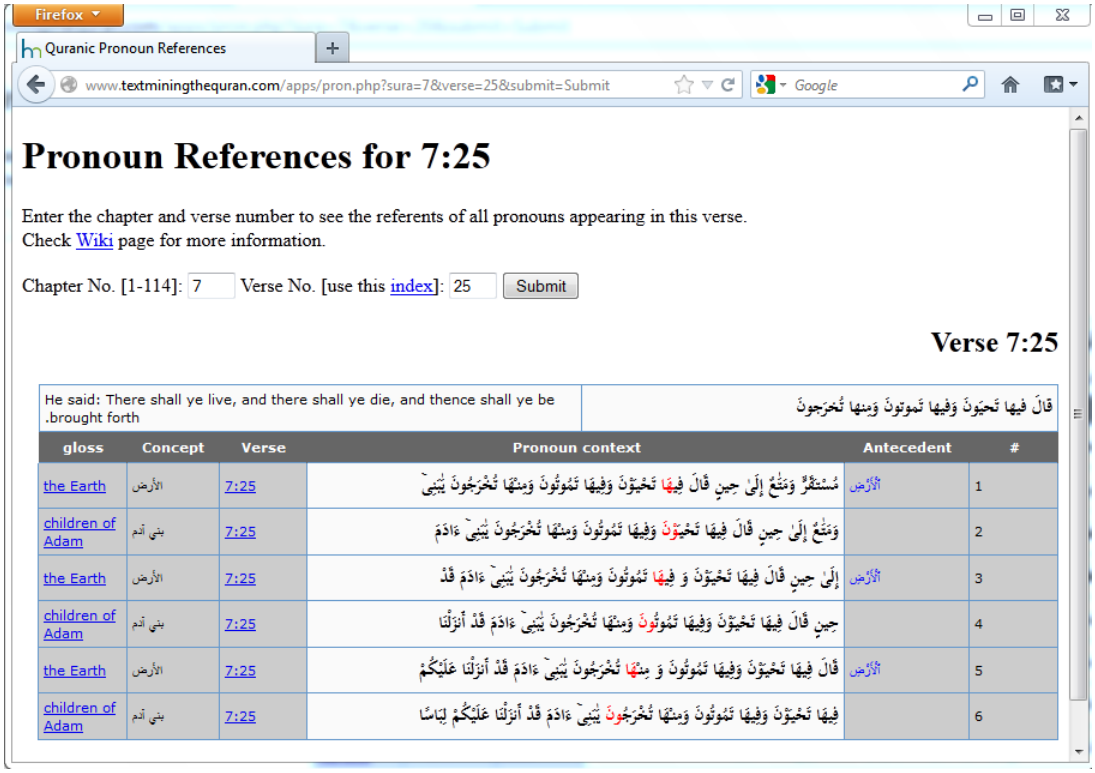


Figure 7.3 - Pronoun resolution of verse 7:25

The user may explore from the concepts listed for this verse, to all other verses that share this same concept, represented as concordance lines for convenient analysis. Figure 7.4 shows instances where the concept ‘children of Adam’ is repeated in the Qur’an as pronouns.

This tool is supposed to be very helpful for Qur’anic researchers, as our QurAna is first of its kind for tagging Qur’anic pronoun referents and allowing to drill down all occurrences of a referent as concordance lines. For example, no existing application can enlist all pronouns in the Qur’an that refers to the ontology entry “Muhammad”. QurAna concept ontology allows browsing all such instances and shows the antecedent word whenever is available. It is noted that “Muhammad” is only mentioned explicitly four times in the Qur’an, but most of the cases where the antecedent is available, the Qur’an uses words like “Messenger” or “Prophet”.

Following is a summary of main features of this application:

- An online wiki page[1] is maintained as documentation of this application.

[1] http://Quran.com

- All instances of pronouns are highlighted in red colour and presented in concordance line style.
- Antecedence string is mentioned whenever is available.
- Reference concept is given both in Arabic and as English gloss.
- Verse number is given and an English translation (Pickthall) is shown.
- The target verse links to a reference Qur'an site[2] for further research.
- English referent gloss is linked with all occurrences of a concept in the Qur'an displayed again in concordance lines style, as depicted in figure 7.4 below.

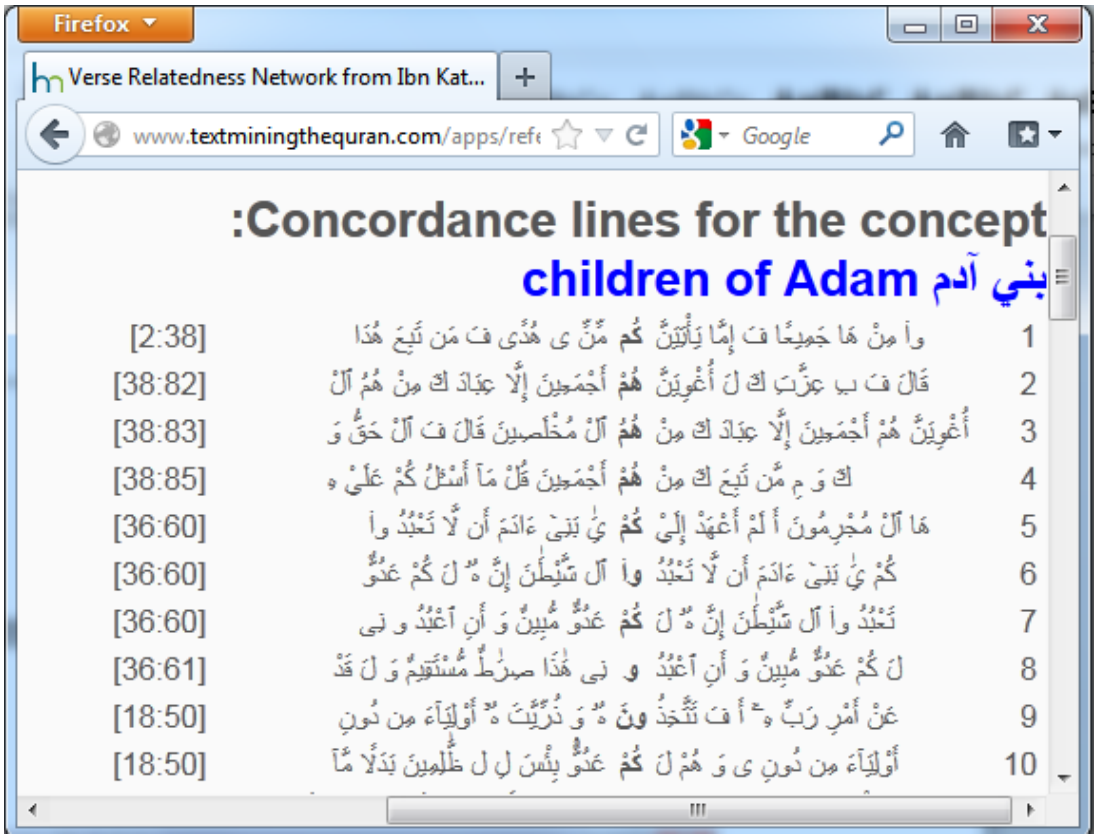


Figure 7.4 - concordance lines for the concept 'children of Adam'

---

[1] Available at http://www.textminingtheQuran.com/wiki/Pronoun_Reference_in_the_Quran

[2] http://Quran.com

7.3 Qur'anic Verse Similarity and Relatedness Application

A set of applications was rendered online to enable researchers to analyze lexical and semantic similarity and relatedness in the Qur'an. In chapter 3.3 a number of examples of verse relatedness were given. Chapter 6 was dedicated to describe our QurSim dataset where nearly 8,000 pairs of related verses were compiled from Tafsir Ibn Katheer.

7.3.1 Lexical Similarity

A number of query pages were rendered online for searching lexical similarity given an input verse.

7.3.1.1 PHP Similar_Text function

This online application[1] adapted PHP Similar_Text function[2] to the Arabic raw text of the Qur'an. This calculates the similarity between two strings as described in (Oliver 1993). The complexity of this algorithm is O(N**3) where N is the length of the longest string. This function takes as input a verse and a threshold value of the percentage of lexical similarity between the two texts. Then the function exhaustively compares similarity between the entered verse and all other verses in the Qur'an, and returns the ones that fall above the threshold percentage.

---

[1] Available online at http://www.textminingtheQuran.com/php/similarity.html

[2] Online manual of this function is available at http://php.net/manual/en/function.similar-text.php

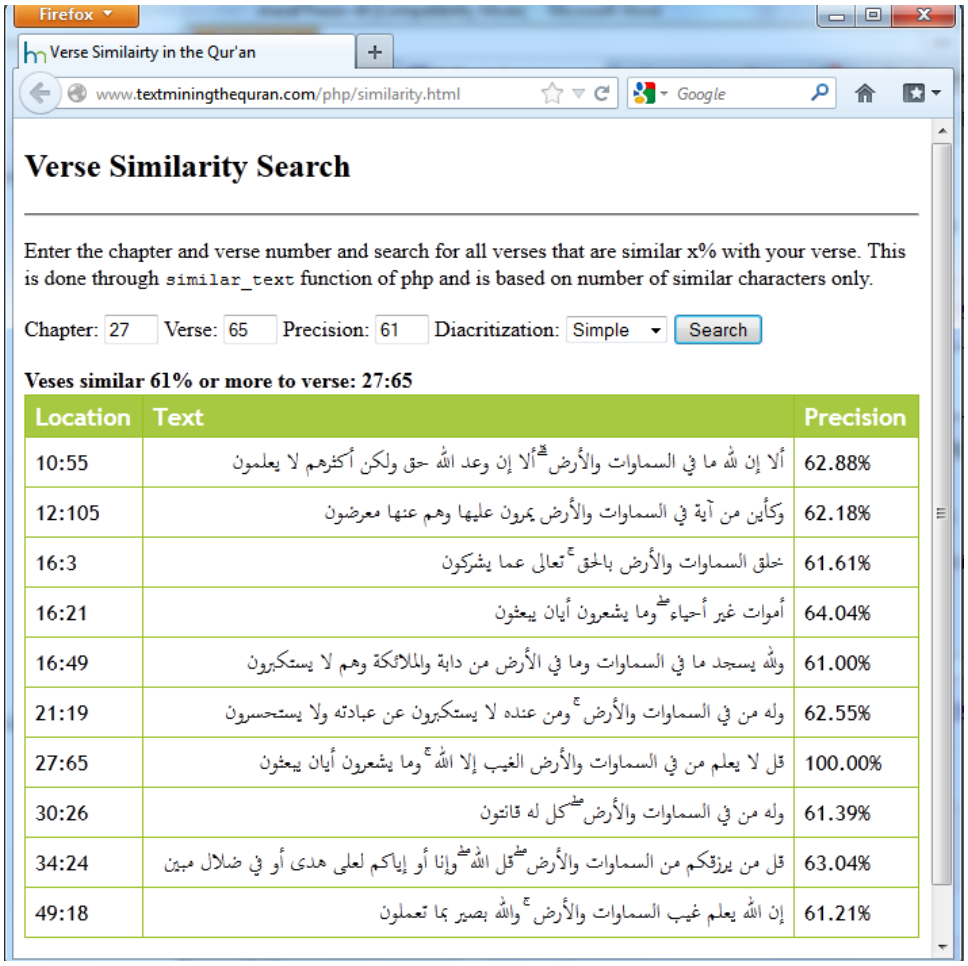


Figure 7.5 – Lexical similarity using PHP `Similar_Text()` function

Figure 7.5 shows the query page of verses similar by at least 60% to the verse 27:65. The search also allows comparing verse similarity against the default un-vowelized or vowelized text.

7.3.1.2 Text::Similarity::Overlaps Module

This is a Perl module[1] that measures the similarity of two files or two strings based on the number of overlapping (shared) words, scaled by the lengths of the files. It computes the F-Measure, the Dice Coefficient, the Cosine, and the Lesk measure (Lesk 1986).

This module was adapted into an online query page[2] to measure similarity between Qur'anic verses, and allows users to input verse number and specify a percentage similarity level. Figure 7.6 shows the screenshot of the system again

[1] Available at http://www.d.umn.edu/~tpederse/text-similarity.html

[2] Available online at http://textminingtheQuran.com/cgi/similarity.html

taking verse 27:65 as a sample to help compare the two lexical similarity measures.

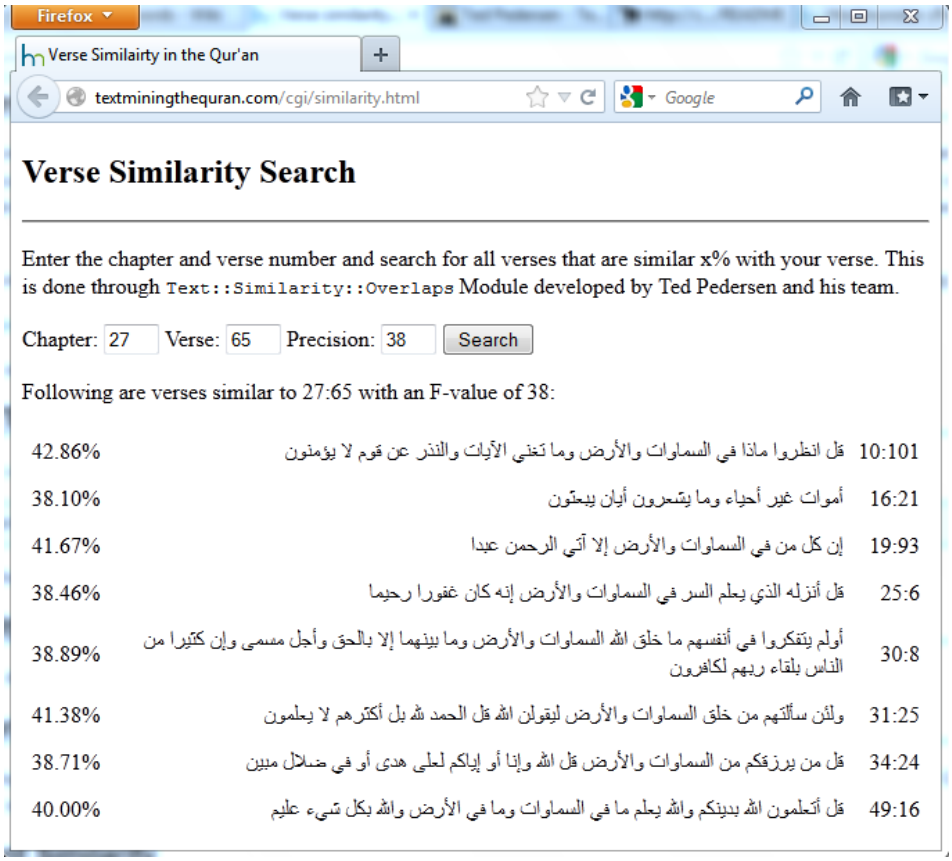


Figure 7.6 – lexical similarity using `Text::Similarity::Overlaps` module

Both Perl and PHP functions work at lexical level and precision level can be adjusted. I noticed from running the same input verse (i.e., 27:65) on both application that precision level varied greatly between the two. While PHP `similar_text` function returned 10 verses when setting precision level to 61, the perl `text::similarity::overlap` module returns zero results at that precision level, and we get only 8 returned verses when precision is set to 38. Also, comparing the output of both models (i.e., 10 verse from PHP and 8 from Perl), we have only 2 common verses returned. This discrepancy shows that lexical methods alone are not very helpful for automatic detection of relatedness and similarity between short text, if not supplemented with more semantic treatment. This finding motivated us to collect the QurSim dataset.

7.3.1.3 Vector Similarity using TF-IDF

The vector space model is widely used in information retrieval where the distance between the query terms and each document, represented as vectors, is measured by comparing the cosine of the angle between the vectors (Manning et al 2008). I followed the same methodology and considered each verse of the

Qur'an as a separate document. The first step is to consider taking out stop words. I resorted to QAC tags to define the list of stop words, which included all types of Qur'anic particles as well as pronouns (including demonstrative and relative pronouns) and adverbs[1]. Our stop-words list of the Qur'an takes out 33,931 words keeping only 43,499 core words in the Qur'an. So each verse of the Qur'an is reduced into content words only. As some verses are very short (like those one-word verse that contain a single Qur'anic initials), there are cases when an entire verse is excluded. Next for each remaining content word TF-IDF value was assigned according to the formula:

$$tf\text{-}idf = tf \times idf$$

where,

$idf = \log (N/df)$ and tf is the count of the word normalized by the total length of the verse. N is the total number of documents which in our case is the total number of verses which have at least one content word. This amounted to 6,214 verses. (note that in total Qur'an has 6,236 verses, so we had 22 short verses having no content words. df is the document frequency, i.e., how many verses a particular Qur'anic root is appearing.

Given the above values, an online query page[2] was created where a user inputs a verse, and the application compares the dot product of this verse against all tf-idf values of other verses and returns the 10 most highest values. Figure 7.7 again returns the verse similar to 27:65.

---

[1] See all QAC tagset at http://corpus.quran.com/documentation/tagset.jsp

[2] Available at http://www.textminingtheQuran.com/apps/tfidf.php

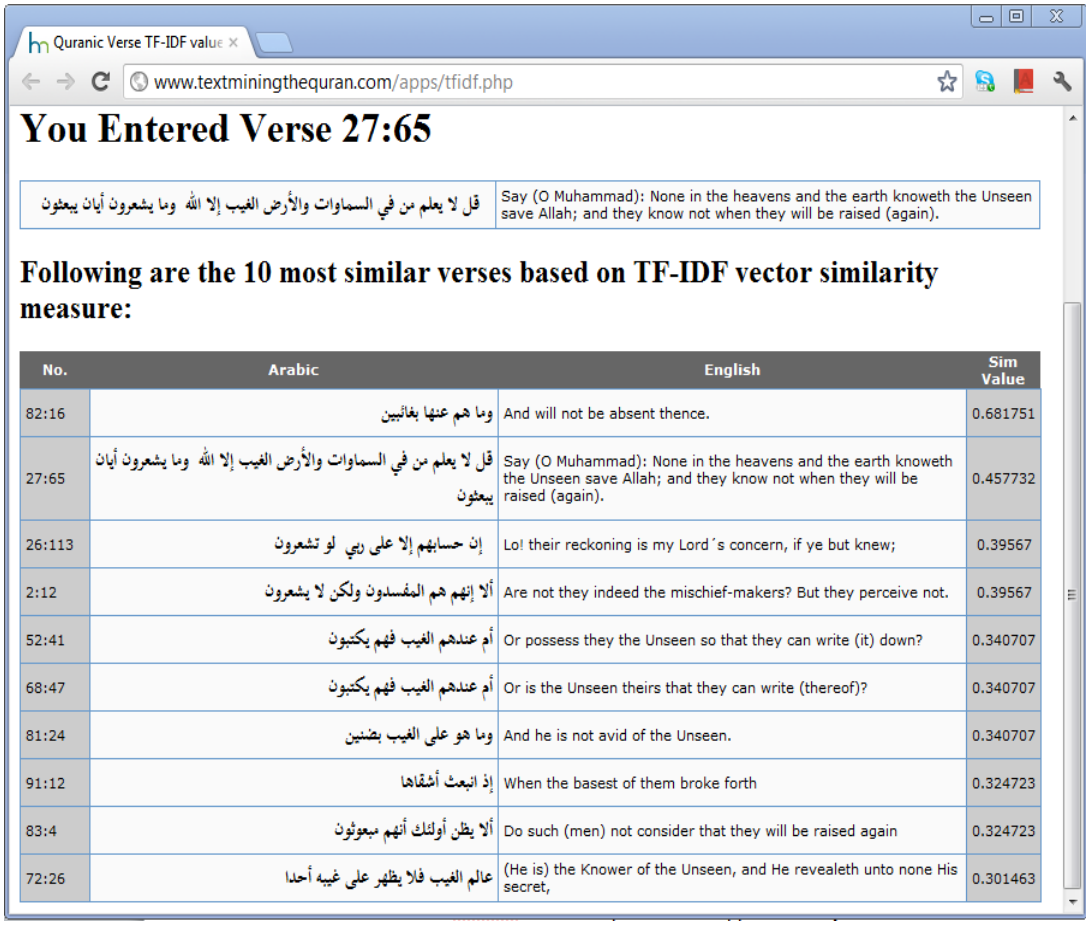


Figure 7.7 – lexical similarity using vector similarity (TF*IDF)

It has been noted that the results does not seem to be very useful, as the results are biased toward similar length verses which has common keywords shared in between. Shared keywords do not necessarily make two verses related especially when these shared keywords are very common Qur'anic words like 'say', 'do', etc.

7.3.1.4 Verse segment lexical similarity

So far our lexical similarity measures considered a verse as the smallest unit for similarity comparison. However, as we have seen, Qur'anic verses vary greatly in size from being a single word (e.g., verse 20:1), to a full page verse of 128 words (which is verse 2:282). The Qur'anic verses has been marked by early scholars into segments separated by pause and prosodic markers. Following is an illustrative example from verse 2:255 where one single verse has been marked into 9 smaller segments. (translatios taken from 'Saheeh International Edition')

Allah! There is no deity save Him, the Alive, the Eternal [MARKER]	اللَّهُ لَا إِلَٰهَ إِلَّا هُوَ الْحَيُّ الْقَيُّومُ ۚ

Neither slumber nor sleep overtaketh Him. [MARKER]	لَا تَأْخُذُهُ سِنَةٌ وَلَا نَوْمٌ ۚ
Unto Him belongeth whatsoever is in the heavens and whatsoever is in the earth. [MARKER]	لَّهُ مَا فِي السَّمَاوَاتِ وَمَا فِي الْأَرْضِ ۗ
Who is he that intercedeth with Him save by His leave? [MARKER]	مَن ذَا الَّذِي يَشْفَعُ عِندَهُ إِلَّا بِإِذْنِهِ ۚ
He knoweth that which is in front of them and that which is behind them, [MARKER]	يَعْلَمُ مَا بَيْنَ أَيْدِيهِمْ وَمَا خَلْفَهُمْ ۖ
while they encompass nothing of His knowledge save what He will. [MARKER]	وَلَا يُحِيطُونَ بِشَيْءٍ مِّنْ عِلْمِهِ إِلَّا بِمَا شَاءَ ۚ
His throne includeth the heavens and the earth, [MARKER]	وَسِعَ كُرْسِيُّهُ السَّمَاوَاتِ وَالْأَرْضَ ۖ
and He is never weary of preserving them.. [MARKER]	وَلَا يَئُودُهُ حِفْظُهُمَا ۚ
He is the Sublime, the Tremendous. [END]	وَهُوَ الْعَلِيُّ الْعَظِيمُ

Qur'anic Verse - 2:255

Given the above significance of verse segmentation, I have modified the text_similarity method to compare verses at segment level[1]. The results show better performance when shorter segments are considered. Table 7.1 lists the closest verses returned against the verse 27:65 using the above mentioned lexical similarity measures.

`Text_similarity` at segment level	`Text_similarity` at verse level	`text::similarity`	Tf-idf
10:18	10:55	10:101	82:16
6:12	12:105	16:21	27:65
10:101	16:3	19:93	26:113
13:16	16:21	25:6	2:12
16:21	16:49	30:8	52:41
19:93	21:19	31:25	68:47
21:19	30:26	34:24	81:24

[1] Available online at http://textminingtheQuran.com/cgi/sim-phrase.html

Table 7.1 – comparison of different similarity measures against verse 27:65

This comparison table shows that TF-IDF measure performs the worst. It is not surprising though that TF-IDF method performed worst. The size of the Qur'an is smaller compared to standard corpora. This results in a particular term appearing only few times creating a very sparse matrix lowering the value of dot products when measuring cosine similarity between two verses. Moreover, TF-IDF considers only lexical similarity while the Qur'an observes high level of semantic similarity, for example, the word 'Qur'an' have been repeatedly interchanged with words meaning 'book', 'criterion', 'light', etc.

7.3.2 Semantic Relatedness through QurSim

In previous subsection we noticed that all lexical comparison methods fail to relate verses that are semantically similar, but has no common root or keyword. This was the motivation behind collecting the QurSim dataset. Refer to Chapter 6 for detailed discussion on the design and evolution of QurSim. Here I present an application layer using QurSim dataset. This application enables performing queries on our developed QurSim dataset. This application is rendered online[1] through PHP scripts with MySQL as the backend database. A detailed documentation page is maintained online as well[2].

Upon entering an input verse by a user, the following information is returned by this application.

7.3.2.1 Graph of network of verses

For better visualization of related verses, the Dracula Graph Visualization tool[3] was used. The graph depicts Qur'anic verses as ovals and arrows shows the relations between verses. The label over arrow shows the number of common words between a pair of verse. The application goes one level deeper and brings the relations of the related verses as well. In this way, the user can traverse to two level of relativeness in a visually appealing format. The graph allows users to re-position the verse nodes for better visualization. Figure 7.8 shows the network of related verses for the verse 27:65

---

[1] Available online at http://www.textminingtheQuran.com/apps/similarity.php

[2] Can be accessed at http://www.textminingtheQuran.com/wiki/Verse_relatedness_in_Ibn_Kathir

[3] Available at http://www.graphdracula.net

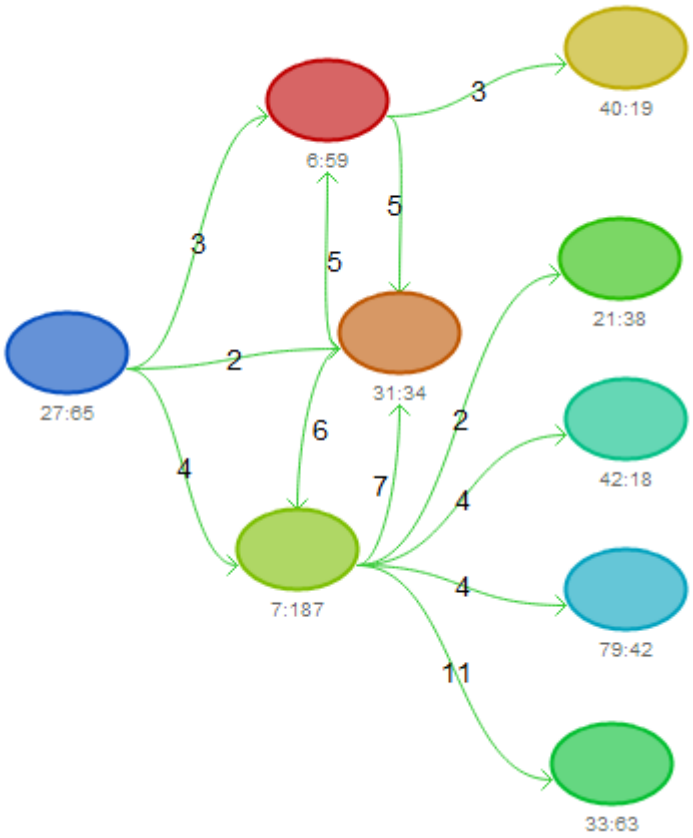


Figure 7.8 – Network of verses related directly or indirectly to verse 27:65 from QurSim dataset

7.3.2.2 directly and indirectly related verses

The application returns a tabular list of those verses that are directly linked to the input verse from QurSim dataset. Directly linked means those pairs that are branded by Ibn Katheer as related. Figure 7.9 shows the result for verse 27:65.

Then, indirectly related verses are automatically generated from the database, by recursively visiting all direct relations of the directly related verses that are already retrieved. This is considered a novel extension to the Ibn Katheer list, and is thought to be interesting to the Qur'anic researchers whereby they will be able to construct a cluster of verses surrounding the main topic of the original input verse. For example, the main topic of verse 27:65 is to inform that only Allah knows the unseen and future knowledge. Surrounding this topic and using QurSim, one is able to bring eight relevant verses with a one-click, whereas investigating books of Tafsir would require spending considerable amount of time

retrieving related verses from multiple locations. However, it is not that transitivity property is not always true and not all indirectly traversed verses show relatedness with the original verse. As this list is generated for human consumption, he or she should easily figure out the related from unrelated verses.

You Entered Verse 27:65

قُل لا يَعلَمُ مَن فِي السَّمٰوٰتِ وَالأَرضِ الغَيبَ إِلَّا اللَّهُ وَما يَشعُرونَ أَيّانَ يُبعَثونَ	Say (O Muhammad): None in the heavens and the earth knoweth the Unseen save Allah; and they know not when they will be raised (again). **Pronoun Referents:** the polytheists,

Following are 3 verses directly related to 27:65 from Ibn Kathir:

No.	Arabic	English	Common Roots	Level
7:187	يَسـَٔلونَكَ عَنِ السّاعَةِ أَيّانَ مُرسىٰها قُل إِنَّما عِلمُها عِندَ رَبّي لا يُجَلّيها لِوَقتِها إِلّا هُوَ ثَقُلَت فِي السَّمٰوٰتِ وَالأَرضِ لا تَأتيكُم إِلّا بَغتَةً يَسـَٔلونَكَ كَأَنَّكَ حَفِيٌّ عَنها قُل إِنَّما عِلمُها عِندَ اللَّهِ وَلٰكِنَّ أَكثَرَ النّاسِ لا يَعلَمونَ	They ask thee of the (destined) Hour, when will it come to port. Say: Knowledge thereof is with my Lord only. He alone will manifest it at its proper time. It is heavy in the heavens and the earth. It cometh not to you save unawares. They question thee as if thou couldst be well informed thereof. Say: Knowledge thereof is with Allah only, but most of mankind know not. **Pronoun Referents:** the infidels of Quraish, Prophet Muhammad, the Hour, Allah, mankind,	4	2
6:59	وَعِندَهُ مَفاتِحُ الغَيبِ لا يَعلَمُها إِلّا هُوَ وَيَعلَمُ ما فِي البَرِّ وَالبَحرِ وَما تَسقُطُ مِن وَرَقَةٍ إِلّا يَعلَمُها وَلا حَبَّةٍ فى ظُلُمٰتِ الأَرضِ وَلا رَطبٍ وَلا يابِسٍ إِلّا فى كِتٰبٍ مُبينٍ	And with Him are the keys of the Invisible. None but He knoweth them. And He knoweth what is in the land and the sea. Not a leaf falleth but He knoweth it, not a grain amid the darkness of the earth, naught of wet or dry but (it is noted) in a clear record. **Pronoun Referents:** Allah, keys of the unseen, tree leave,	3	2
31:34	إِنَّ اللَّهَ عِندَهُ عِلمُ السّاعَةِ وَيُنَزِّلُ الغَيثَ وَيَعلَمُ ما فِي الأَرحامِ وَما تَدرى نَفسٌ ماذا تَكسِبُ غَدًا وَما تَدرى نَفسٌ بِأَىِّ أَرضٍ تَموتُ إِنَّ اللَّهَ عَليمٌ خَبيرٌ	Lo! Allah! With Him is knowledge of the Hour. He sendeth down the rain, and knoweth that which is in the wombs. No soul knoweth what it will earn to-morrow, and no soul knoweth in what land it will die. Lo! Allah is Knower, Aware. **Pronoun Referents:** Allah,	2	2

Figure 7.9 – verses directly related to verse 27:65 from QurSim

Following are the various information returned for each related verse:

- Arabic (from Tanzil project) and English translation (Pickthall translation) for the input verse.
- Arabic (from Tanzil project) and English translation (Pickthall translation) for each of the directly related verses.
- A count of number of common root words between the related verse pairs. This count was made through counting common roots using QAC which specifies root word for each Qur'anic word.
- Each related verse can be investigated further to its original context in Ibn Takhir tafsir book through a hyperlink. This allows the user to check the context of relationship in the commentary book.

- Each related verse pair is given a relevance indicator showing the level of relatedness –either strong or weak. See section 6.3.1 for further details on these levels.
- Each verse is integrated with QurAna dataset and pronoun referents are shown.

7.3.2.3 concept clouds

Because we have access to the concepts referred to by the pronouns of any verse from our QurAna dataset, we can easily generate a concept cloud from all verses related –directly or indirectly- to our input verse. This gives the reader an overview of the main theme of all the verses. Figure 7.10 below shows the concept cloud from all verses related to 27:65.

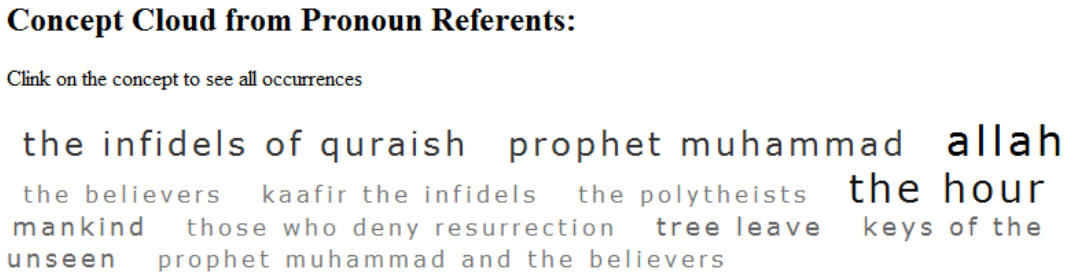


Figure 7.10 – Concept cloud from pronouns of all verses related to verse 27:65

7.3.3 Semantic relations between Qur'anic chapters

Domain experts can utilize the QurSim dataset for more interesting investigations in Qur'anic studies. For example, a Qur'anic student might want to find relatedness between Chapters rather than verses. Using QurSim, we can relate two chapters by the frequency of cross-reference between their verses. Figure 7.11 is an example from our online application[1] that shows such relations. Nodes in this graph show chapters and the number associated with arrows show the number of cross-referenced verses between the chapters. Qur'anic chapters are broadly categorized thematically into Meccan or Medinan chapters distinguished in our graph as red or green respectively. See chapter 8 for a detailed treatment of this classification and its significance. From figure 7.11, for example, a

[1] Available at http://textminingtheQuran.com/apps/surah.php

Qur'anic student realizes that although, chapter no. 2 'al-Baqarah' is Medinan, still it has tight relation with chapter 7 'al-A'raf' especially those verses related with the story of Moses. Also, one could see the close relation between chapter 2 and chapter 4 'an-Nisa'' based on many common verses related to ruling of marriage and women related matters. (in fact, 'an-Nisa' in Arabic means 'women').

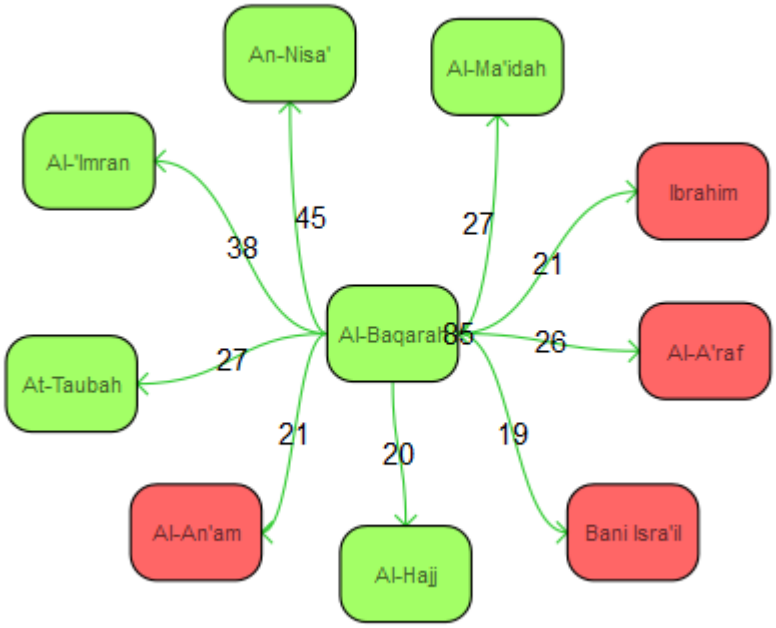


Figure 7.11 - Relatives of chapter No. 2 "Al-Baqarah", red nodes are Meccan chapters, whereas green nodes represent Medinan chapters

7.4 n-Gram Search

N-gram search takes an input word and returns in descending order of frequency all Qur'anic phrases containing this word. For example, figure 7.12 below shows the 5 most n-gram terms returned for the word "الله" Allah, which is the most frequent word in the Qur'an. Like other applications, nGram is rendered into online web page[1], and the search word can be included in the URL.

[1] Available at http://www.textminingtheQuran.com/apps/ngrams.php

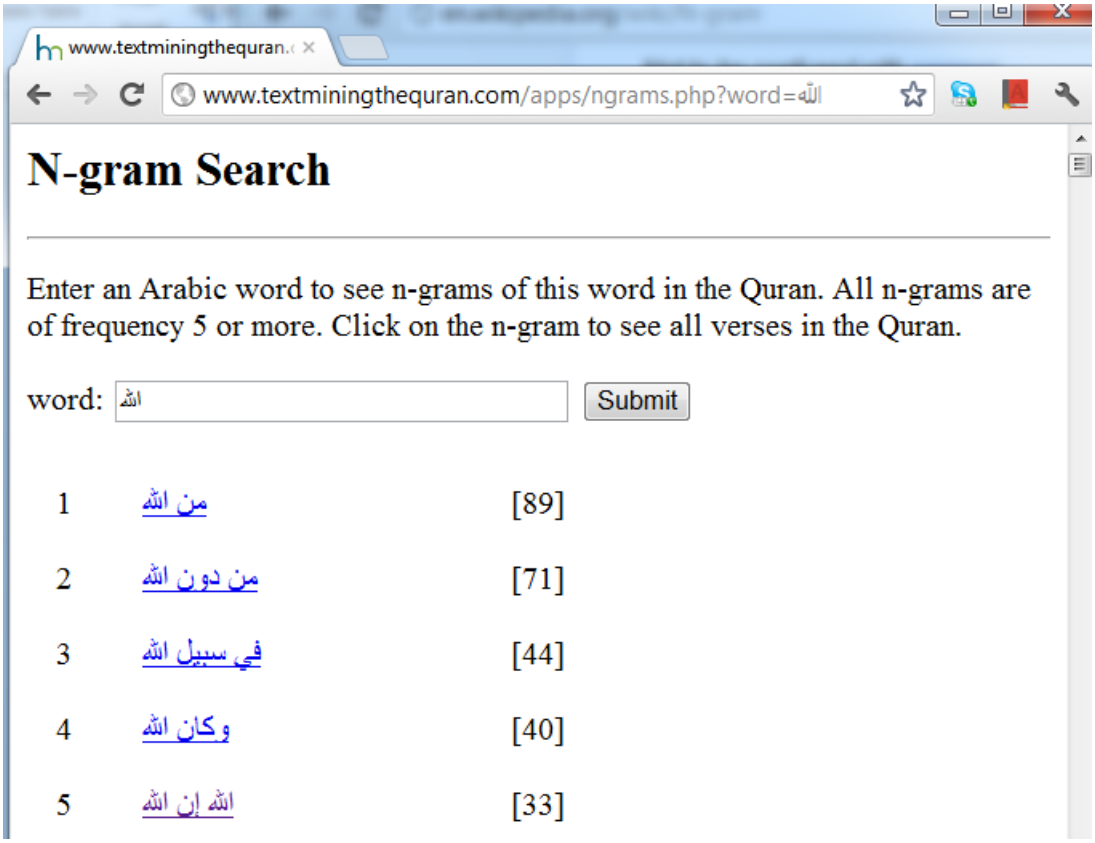


Figure 7.12 – The top 5 n-gram patterns for the word Allah with their frequencies.

This tool is helpful for Qur'anic researchers to discover the frequent patterns of a given word. "Allah" being the most frequent and central word in the Qur'an, a researcher might wonder in what linguistic context this word appears most in the Qur'an. He/she realizes thanks to this tool that the construct (من الله) "from Allah" is the most frequent patter, followed by (من دون الله) "other than Allah" and then (في سبيل الله) "in the cause of Allah". All instances can be visited as concordance lines by clicking on the link. Figure 7.13 below are the first 5 instances for the second most frequent n-gram for the word Allah, which is "other than Allah".

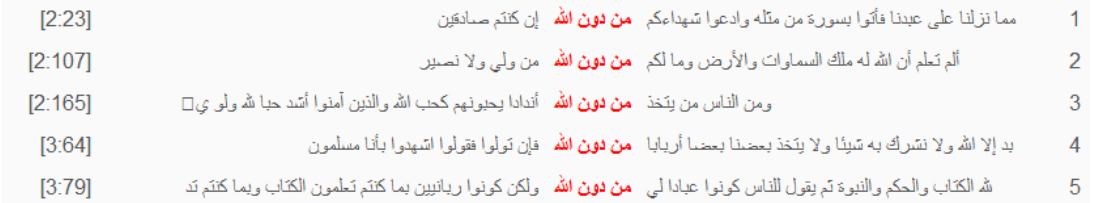
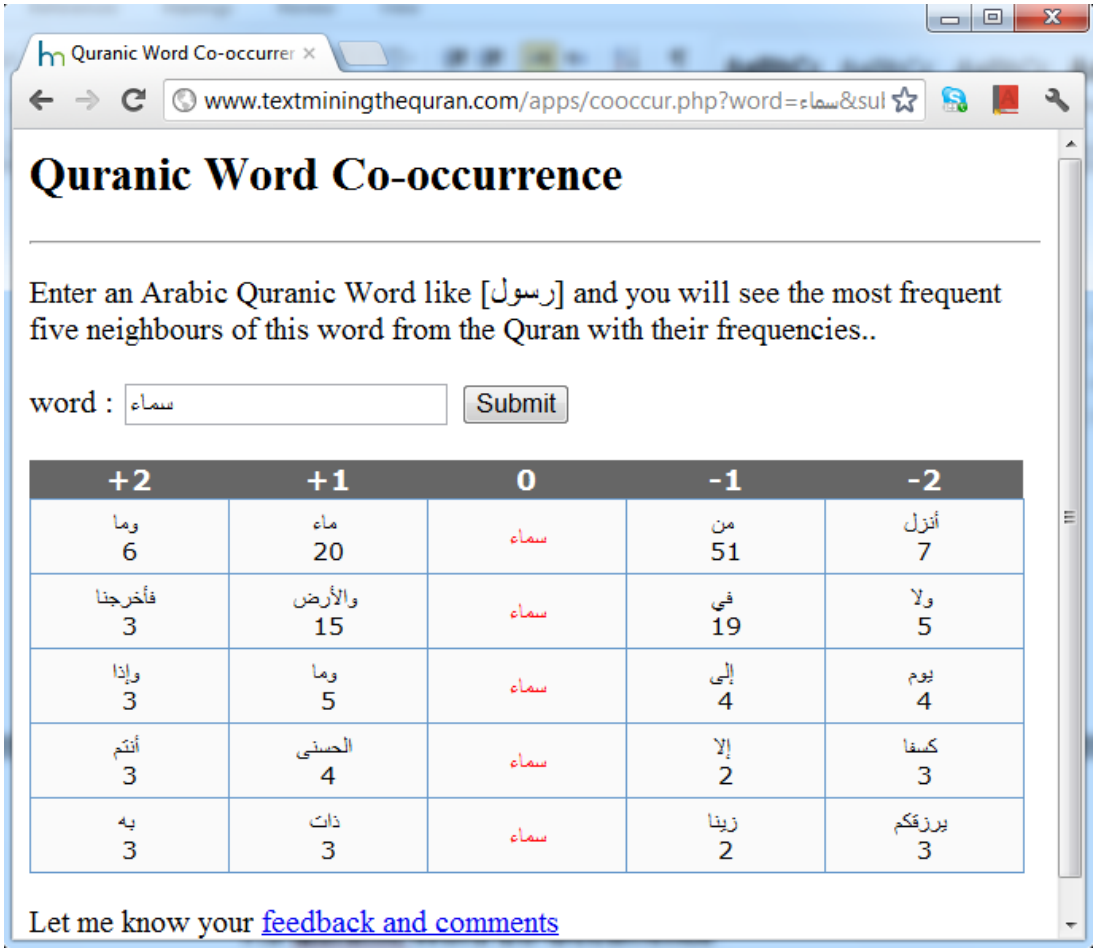


Figure 7.13 – Top 5 results from the n-gram (من دون الله) "other than Allah"

7.5 Qur'anic Word Co-Occurrence

This online application[1] takes a word and returns the most frequent four neighbors –two words before and two words after- of this word in the Qur'an along with their frequencies.



+2	+1	0	-1	-2
وما 6	ماء 20	سماء	من 51	أنزل 7
فأخرجنا 3	والأرض 15	سماء	في 19	ولا 5
وإذا 3	وما 5	سماء	إلى 4	يوم 4
أنتم 3	الحسنى 4	سماء	إلا 2	كسفا 3
به 3	ذات 3	سماء	زينا 2	يرزقكم 3

Figure 7.14 – Co-occurrences of the Qur'anic word (سماء) "sky/heaven"

For example, figure 7.14 above gives the result for the word (سماء) meaning "sky" or "heaven". We notice that the most frequent immediate word before this word is the preposition (من) or "from" repeated 51 times, and the immediate word after is (ماء) "water" repeated 20 times, as in the phrase "Allah sends down from sky water". This application would be very helpful for stylistic and statistical language study of the Qur'an.

[1] Available at http://www.textminingtheQuran.com/apps/cooccur.php

7.6 Color coded POS display

This online application[1] relies on the POS tagging of QAC. It displays four main POS tags: noun, pronoun, verb and adjective, each in different color as depicted in figure 7.15.

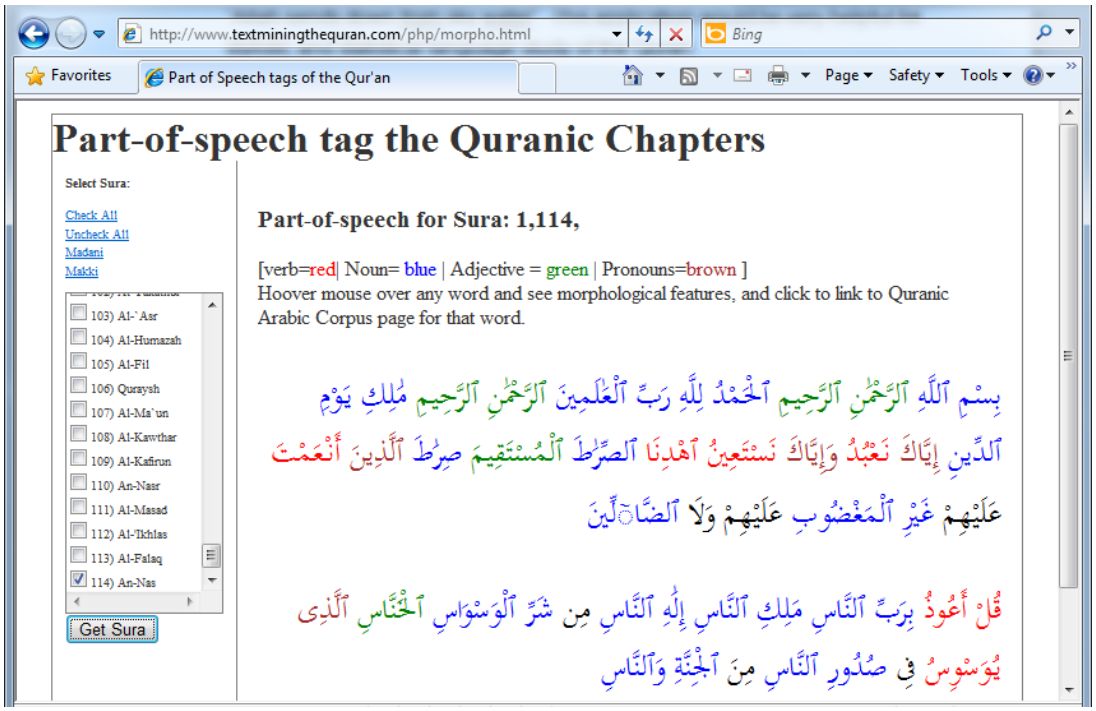


Figure 7.15 – color coded POS display of chapters 1 and 114

Note that the application allows selecting multiple chapters at a time, and upon pointing the mouse over any word a detailed morphological features are displayed as tooltip box. Also, as a special selection, all Meccan and Medinan chapters can be selected through pre-defined list.

This kind of simple display application helps visualize a chapter based on POS. For example, it was clear from the above example that ‘noun’ (colored in blue) is the most used POS for both chapters 1 and 114.

7.7 QurCloud: Qur’anic Chapter Word Cloud

This online application[2] enables users to select one or more chapters from the Qur’an and will display ‘word cloud’ for the selected chapters. The significance of

---

[1] Available at http://www.textminingtheQuran.com/php/morpho.html

[2] Available at http://www.textminingtheQuran.com/php/wordcloud.html

a word –based on its frequency- is directly related with the size of the display font of this word. Most prominent words are displayed in bigger fonts dictating an overarching theme for a particular chapter. For example, figure 7.16 below is a word cloud for the last two chapters of the Qur'an, i.e., 113 and 114.

Figure 7.16 – Word cloud from chapters 113 and 114

Following are some features of our QurCloud application.

- Allows selection of multiple chapters
- Allows selection of thematic Meccan or Medinan chapters
- Allows inclusion or exclusion of a list of stopwords[1].
- A wiki documentation page is maintained[2].

[1] Available at http://www.textminingtheQuran.com/php/stoplist.php

[2] Available at http://textminingtheQuran.com/wiki/Word_Cloud

- Word Cloud features are adapted from PHP Tag Cloud scripts[1] developed by 'lotsofcode'.

To the best of our knowledge QurCloud is the first tool of its kind customized for Qur'anic chapters. It will help researchers interested in thematic focus of certain Qur'anic chapter or a group of chapters.

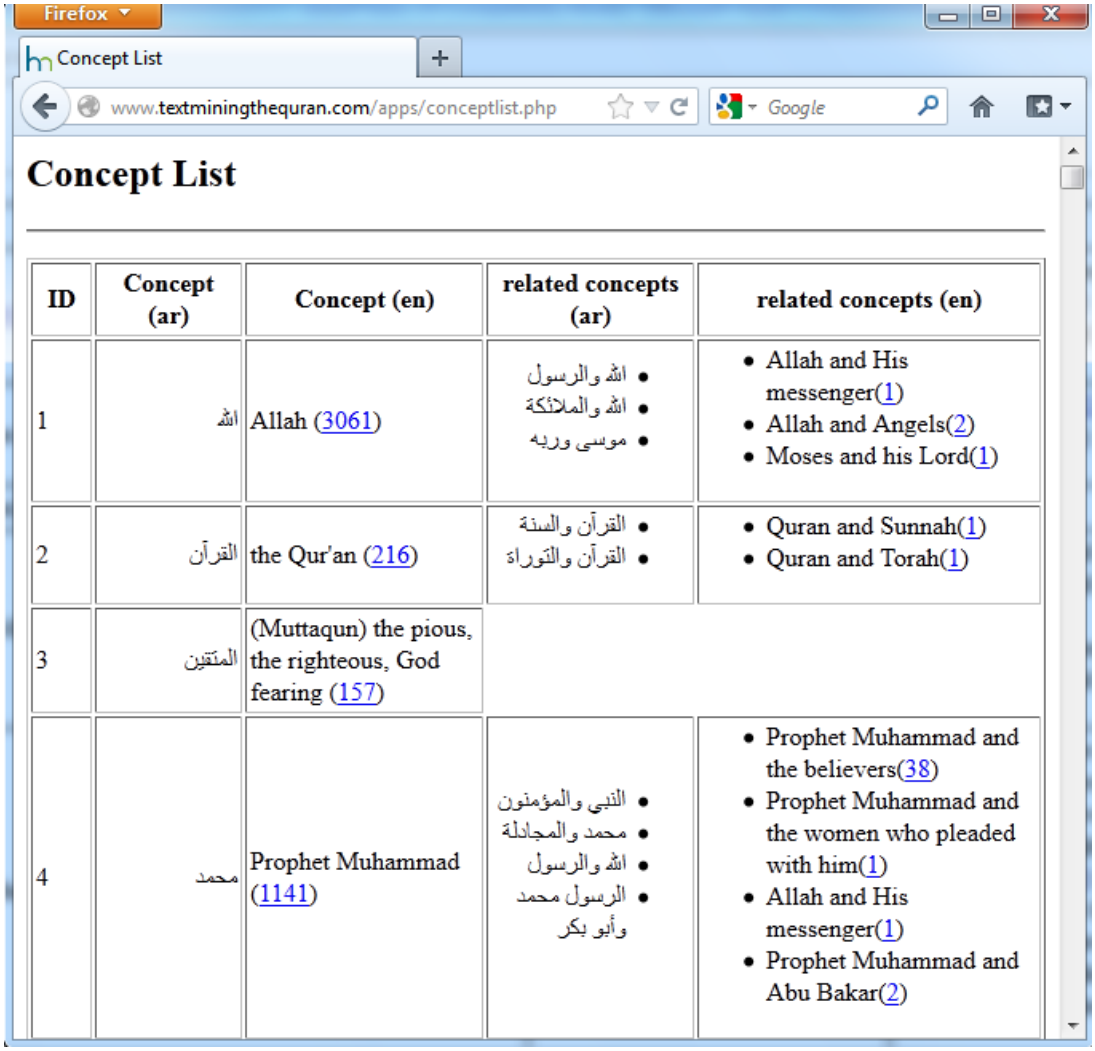
QurCloud can be enhanced in a number of ways through the inclusion of specialized color codes and more option is on controlling the size and font of the display.

7.8 Concept ontology of the Qur'an

As I continued to tag Qur'anic pronouns against the concepts they refer to, I was able to accumulate over 1,000 Qur'anic concepts. An online application was then created to allow users to navigate to the verses attached to a certain concept and find the relationship between concepts. Figure 7.17 is a screenshot of the application where the first four concepts are depicted. Each concept is displayed with (i) Arabic name (ii) English gloss (iii) number of verses carrying this concept (iv) clickable link to a concordance line display of all the verses where this concept appears as a pronoun referent (v) possible relatives of this concept.

This concept map should be very useful text mining tool at the disposal of Qur'anic students. For example, from figure 7.17 below, a student realizes from available Qur'anic search tools, that Prophet Muhammad is referred in the Qur'an only 4 times, however, he is referred to as a pronoun around 1,141 times. Moreover, Prophet Muhammad is associated with the group of believers in the Qur'an as a plural pronoun in over 38 verses, and he/she can retrieve all these instances.

[1] Available at http://hub.lotsofcode.com/tag-cloud/



Concept List

ID	Concept (ar)	Concept (en)	related concepts (ar)	related concepts (en)
1	الله	Allah (3061)	• الله والرسول • الله والملائكة • موسى وربه	• Allah and His messenger(1) • Allah and Angels(2) • Moses and his Lord(1)
2	القرآن	the Qur'an (216)	• القرآن والسنة • القرآن والتوراة	• Quran and Sunnah(1) • Quran and Torah(1)
3	المتقين	(Muttaqun) the pious, the righteous, God fearing (157)		
4	محمد	Prophet Muhammad (1141)	• النبي والمؤمنون • محمد والمجادلة • الله والرسول • الرسول محمد وأبو بكر	• Prophet Muhammad and the believers(38) • Prophet Muhammad and the women who pleaded with him(1) • Allah and His messenger(1) • Prophet Muhammad and Abu Bakar(2)

Figure 7.17 – first few concepts from Qur'anic pronouns. The number refers to the count of verses under each concept.

7.9 Summary

This chapter described some text mining applications made online for querying the Qur'an. These applications were made possible because of the development of various language resources and annotation layers of the Qur'an including QAC, QurAna and QurSim.

Among the text mining applications described in this chapter are: Qur'anic concordancer, QurAna searching application, lexical similarity measures of Qur'anic verses, QurSim query and visualization, Qur'anic concept clouds, semantic relations between Qur'anic chapters, n-gram search of the Qur'an, Qur'anic word co-occurrence, part-of-speech visualization of the Qur'an, Qur'anic chapter word cloud, and concept ontology from pronoun referents.

Chapter 8
Machine Learning Experiments

8.1 Introduction

Machine Learning allows computers to find patterns and predict classifications, clusters, associations and linking threads in a given set of data. A common Machine Learning task is classification. This involves the computer learning how to classify a new example given certain attributes of known examples. To accomplish this task, first human experts need to feed this algorithm with a set of solved examples. The algorithm 'learns' from these examples and predicts the classification of a new, previously-unseen instance. The performance of this algorithm is evaluated by human experts who judge the accuracy of the outcome. We want to define a rich feature set for Meccan and Medinan chapters, and feed the learning program with examples of feature-values for chapters whose classification is "known" and widely agreed, and then ask it to predict the nature of some chapters whose classification is debated by scholars.

The core of the Machine Learning experiment is the extraction of a feature set for Meccan and Medinan chapters; hence, we will first discuss these features. Then, I will describe how these features are encoded in a WEKA (see section 8.3) representation for Machine Learning. Next, I will run the experiment and interpret the results. I will illustrate the visualization tools in WEKA for exploring the data and results. Finally, I will conclude by discussing opportunities for Artificial Intelligence and Machine Learning techniques for other research in Qur'anic studies.

8.2 Meccan and Medinan Chapters

8.2.1 Significance

Studying Meccan and Medinan chapters attracted us for two reasons: one from Qur'anic studies perspective and another from Machine Learning perspective:

(i) **Qur'anic studies perspective**: Qur'anic scholars have shown interest in classifying Qur'anic verses into Meccan and Medinan for many reasons:

(a) It helps in specifying the abrogated and the abrogating verses, as a Meccan verses can be abrogated by a Medinan verse.

(b) It helps in understanding better a chapter or verse by knowing its historical context.

(c) It helps in analysing the evolution and wisdom behind some Islamic rulings in the Qur'an. For example, verse 2:219 acknowledges some benefit in wine while verse 5:90 gives clear prohibition, indicating that chapter no. 2 –although Medinan- revealed before chapter no. 5.

يَسْأَلُونَكَ عَنِ الْخَمْرِ وَالْمَيْسِرِ ۖ قُلْ فِيهِمَا إِثْمٌ كَبِيرٌ وَمَنَافِعُ لِلنَّاسِ وَإِثْمُهُمَا أَكْبَرُ مِن نَّفْعِهِمَا ۗ

They question thee about strong drink and games of chance. Say: In both is great sin, and (some) utility for men;

Qur'anic Verse – 2:219

يَا أَيُّهَا الَّذِينَ آمَنُوا إِنَّمَا الْخَمْرُ وَالْمَيْسِرُ وَالْأَنصَابُ وَالْأَزْلَامُ رِجْسٌ مِّنْ عَمَلِ الشَّيْطَانِ فَاجْتَنِبُوهُ لَعَلَّكُمْ تُفْلِحُونَ

O ye who believe! Strong drink and games of chance and idols and divining arrows are only an infamy of Satan's handiwork. Leave it aside in order that ye may succeed.

Qur'anic Verse – 5:90

(d) Extracting the biography of the Prophet Muhammad and his mission by analysing the chronology of verses that addressed him and discussed his mission.

(ii) **Machine Learning perspective**: Machine Learning algorithms expect as input a number of features and a total count of these features available in each chapter of the Qur'an. Extraction of such features for Meccan and Medinan chapters and counting availability of these features in Qur'anic chapters appeared to us a feasible as well as illustrative for our Machine Learning task.

8.2.2 Source of Information

As this thesis' focus is on Machine Learning for Qur'anic studies, critical analysis of Meccan and Medinan chapters is not the purpose of this thesis. Rather, I only took the issue of Meccan and Medinan as a convenient example to illustrate the potential of Machine Learning for Qur'anic studies. Hence, I will rely in the study of Meccan and Medinan chapters on a research paper by Muḥammad Shafaat Rabbani (Rabbani 2012), as it gives some concise characteristics that would suite our Machine Learning purpose. I will also adapt the classification contained in the index of Medina Qur'an published by King Fahd Complex for the printing of

the Holy Qur'an. Table 8.1 gives a listing of Meccan and Medinan chapters based on this reference, and those chapters which are debatable by scholars are marked with an asterisk.

No.	Category	No.	Category	No.	Category	No.	Category	No.	Category	No.	Category
1	K *	21	K	41	K	61	D*	81	K	101	K
2	D	22	D*	42	K	62	D	82	K	102	K
3	D	23	K	43	K	63	D	83	K*	103	K
4	D	24	D	44	K	64	D*	84	K	104	K
5	D	25	K	45	K	65	D	85	K	105	K
6	K	26	K	46	K	66	D	86	K	106	K
7	K	27	K	47	D*	67	K	87	K	107	K
8	D	28	K	48	D	68	K	88	K	108	K
9	D	29	K *	49	D	69	K	89	K*	109	K
10	K	30	K	50	K	70	K	90	K	110	D*
11	K	31	K	51	K	71	K	91	K	111	K
12	K	32	K	52	K	72	K	92	K*	112	K*
13	D *	33	D	53	K	73	K	93	K	113	K*
14	K	34	K	54	K	74	K	94	K	114	K*
15	K	35	K	55	D*	75	K	95	K		
16	K*	36	K	56	K	76	D*	96	K		
17	K	37	K	57	D	77	K	97	K*		
18	K	38	K	58	D	78	K	98	D*		
19	K	39	K	59	D	79	K	99	D*		
20	K	40	K	60	D	80	K*	100	K		

Table 8.1 - Meccan (K) and Medinan (D) sūra index. (*) indicates a debatable case.

We considered a chapter or a verse as Meccan or Medinan based on its chronology of revelation with respect to the migration of the Prophet Muḥammad from Mecca to Medina regardless of the actual geographic location of its revelation. Thus, a verse will still be Medinan if it is revealed in Mecca after the migration of the Prophet Muḥammad.

We assume that the ultimate verdict on branding a chapter or a verse as Meccan or Medinan lies upon the availability of authentic reporting from those who witnessed the revelation of these verses, i.e., the companions of the Prophet Muḥammad. Thus, I respect that their authentic verdict would override any conflicting verdict that might be deduced from general observation. For example, chapter no. 87 is a Meccan chapter according to the majority of the scholars, even though some claim it to be Medinan based on the observation that the verse 87:14-15 '*qad aflaḥa man tazakka, wa dhakar sma rabbihi faṣalla*' [He has

certainly succeeded who purifies himself, And mentions the name of his Lord and prays.] (Q. 87:14-15) refers to the prayer and alms giving (*zakat*) which were legislated only in Medina. According to tafsīr scholars like al-Qurṭubī and al-sīūṭī either these reference are only linguistic, that means the zakat is not the third pillar of Islam known as zakā rather it is purification of the soul (tazkīyat al-nafs), or these could actually refer to the two pillars of Islam which would be obligated later, as -according to them- a verse can refer to a future ruling.

In this experiment I assumed the classification to Meccana and Medinan on chapter level and not on verse level. Although considering granularity at verse level could produce more accurate results, I have avoided that for three reasons:

(i) There is no standard reference for classification at verse level against which we could benchmark when learning our algorithms.

(ii) The default is that if a chaper is Meccan then all its verses should be Meccan as well, and the same is true for Medinan chapters .

(iii) If we would consider our classification at verse level, then most values will be zero when counting the number of features available in each of the 6236 verses of the Qur'an, resulting in 'data sparseness problem' where abundant values of zeros prevent statistical and Machine Learning process. This problem is still evident at chapter level when considering the small chapters towards the end of the Qur'an, but still it is less evident than when dealing with at verse level.

Finally, I would like to make a disclaimer on the accuracy of our algorithm results. Errors can creep into our data from five sources:

(i) Defining the feature set: Our feature set is based on scholarly observations. If however, the observation itself was erroneous, then that could lead to erroneous results.

(ii) Presence of a few Medinan verses in Meccan chapters or vice versa: Because our classification is on chapter level, these few verses would be misleading "noise" for our algorithm. For example chapter no. 20 is Meccan except probably verse no 130 because it gives an indication on the timing of the five prayers which were made obligatory just before the migration of the Prophet Muḥammad. Another example is the Meccan chapter no. 22 where verses 19-21 are Medinan based on authentic hadith in al-Bukhārī that they were revealed in the context of the battle of Badr .

(iii) Reducing feature sets to keywords: As will be shown later, our Machine Learning algorithm expects counted numbers. To achieve that I had to reduce each feature into a countable keyword to search for, and this reduction process can result in some errors. For example, consider the feature that only Medinan chapters contain detailed family rulings concerning marriage, divorce, etc., I reduced this feature into four searchable roots: nkḥ (marry), ṭlq (divorce), nisāᵓ (women) and zaūj (pair, mate, spouse). However, selecting the term zaūj (pair) will cause noise when it is not used in a context of detailed rulings on marriage: such as in the case of description of the women in paradise (e.g. verse. 44:54), or pair of all living things as in verses 26:7 and 31:10, or reference to Adam and his wife (e.g., verses 7:19 and 20:117), or reference to pair of every animal in the story of Noah as in verse 11:40.

(iv) Exceptional Nature of some Qur'anic chapters: Some Medinan chapters exhibit features of Meccan chapters and vice versa. For example, one of our features states that any chapter that tells stories of previous nations and prophets is a Meccan chapter, another feature states that any chapter that tells the story of Adam and ᵓIblīs (devil) is Meccan, however, we notice that chapter no. 2 has these features but still it is a Medinan chapter. These exceptional cases are not captured in our algorithm.

(v) In some cases I have resorted in our morphological search to the Qur'anic Arabic Corpus (QAC) - which is still undergoing manual verification, although to our knowledge the results are accurate- still any mistake in this resource will be reflected in the counting statistics for our Machine Learning algorithms.

8.2.3 Features of Meccan and Medinan Chapters

Following is a list of 14 features I selected in my experiments. These features are based on scholarly observations. A subset from a more exhaustive set is selected here keeping in the mind the counting need of our algorithms, hence, there could be more prominent features but less feasible to incorporate into our algorithms. For example, Meccan chapter is characterized by frequent argumentation with the polytheists of Mecca, but it was not trivial to reduce this feature into searchable and countable keywords, and hence I excluded it from my selected list below:

(1) Verses of *Sajdah* (command for prostration): The chapters which contain such verses are labelled Meccan except chapters no. 13 and 22.

(2) The aversion letter '*kalla'*: Any chapter containing this word is labelled Meccan.

(3) Any chapter containing the phrase '*ya ayyhan nasu'* (O Mankind) but not the phrase '*ya ayyiha alladhyna aamanu'* (O you who believe) is a Meccan chapter. However, chapter no. 22 is an exception where both of these constructs are present. According to (Rabbani 2012), this chapter is Medinan, however, there are debate among scholars on this issue. One thing we can tell for sure is that this chapter has some verses revealed in Meccan period and other verses revealed in Medinan period.

(4) Likewise, starting a verse with '*ya ayyiha alladyna aamanu'* (O you who believe) is a characteristics of Medinan chapter.

(5) Any chapter that starts with initials like '*alim laam meem'* or '*alim laam raa*' is a Meccan, except chapters no. 2, 3 and 13.

(6) Any chapter telling the story of Adam and Iblis (devil) is Meccan except chapter no. 2.

(7) Any chapter telling the stories of previous nations and prophets like Noah, Lut, Aad, *Thamoud* and *Shuayib* is a Meccan sura.

(8) Meccan chapters contain abundant use of linguistic instrument of denunciation, excoriation, emphasis and oath.

(9) The verse length of Meccan chapters are usually shorter than Medinan chapters.

(10) Meccan chapters give more emphasis on eschatological topics.

(11) Medinan chapters describe rulings of Jihad and tell about the great battles like Badr, Uhud, and al-Ahjab.

(12) Medinan chapters detail on the rulings concerning marriage, breastfeeding, divorce and similar family matters.

(13) Medinan chapters are characterized by repeated dialogue and arguments with the people of the Book, i.e., Jews and Christians.

(14) Medinan chapters are characterized by referring to acts of worship in Islam, e.g., the pillars of Islam like prayer, zakat, fasting, and Hajj.

We now turn to translate each of the above features into searchable and countable keywords, but before that, a few words need to be said on our sources for such keyword search.

8.2.4 Searching resources

To cater for searching for these features we need a source that enables search beyond just keywords and includes a search over root words and various kinds of particles. I resorted to four sources for extracting these features as follows:

(i) Al-mu'jam al-mufahras li al-fadh al-Qur'an by Muhammad Foad Abdul Baqi (MMAQ) (Abdulbaqi 1955): this is a popular index manually checked and verified for correctness. This index lists Qur'anic words by alphabetical order of their roots and gives counting for each root's verbs and then nouns. This index however does not include *Huruf* (particles), nor does it allow search by phrases.

(ii) The Qur'anic Arabic Corpus by Kais Dukes (QAC) (Dukes and Habash 2010): This is an online resource recently developed as part of computational research at University of Leeds. This online resource lists a number of morphological features for each Qur'anic word like: gender, person, number, root, prefixes, suffixes, verbs form, voice and mood. It also gives the part-of-speed tag for each word. The lemmas and roots are stored as Buckwalter transliterations. These features are machine generated followed by continuous manual verification.

(iii) Phrase Search at Qur'anComplex.Com (QC): I used the advanced search facility at [http://www.Qur'ancomplex.org/Qur'an/Search/search.asp accessed on 30th June 2012].

(iv) Qur'an Tanzil project (QTP): This is an online resource available at [http://tanzil.info accessed on 30th June 2012]. This site allows downloading a verified plain text of the Qur'an. I shall use this resource as will be explained later to count the average length of the verses of each chapter.

8.2.5 Counting Features

After identifying the features and finding the resources to find these features we now turn to the actual searching and counting of these features. Following is a further consideration of these features with discussion on our reduction criteria of these features into feasible searchable terms as well as a note on possible noise this reduction may cause in terms of erroneous inclusion or exclusion of some valid terms. The sources below are MMAQ unless otherwise mentioned.

(1) Verses of *Sajdah* (command for prostration): I wanted to extend the normal count of verse of *sajdah* to count how many times the root sjd (prostrate) appears in each chapter. This extension is likely to give more counts and thus gives us an opportunity to discover if this results in interesting correlation. With this search I

found 92 occurrences of this root, however that comes with the expense of inclusion of certain Medinan ayat that contains the root sjd, for example masjid (mosque) has this root and as such 14 of the 15 occurrences of al-masjid al-haraam are Medinan (e.g., Q. 2:144, 149, 150).

(2) The aversion letter '*kalla'*: these are found 33 times and confirmed to be Meccan in all cases.

(3) The phrase *ya ayyihan naas* (O Mankind): searching QC for this phrase returns 20 instances half of them are Medinan (2 instances in sura no. 2, 3 in sura no. 4, 4 in sura no. 22, and once in sura no. 49). Given this 50-50 nature of this feature makes it not beneficial as a classification criterion, and hence I will exclude it from our final feature set.

(4) The phrase *'ya ayyiha alladhina aamanu'* (O you who believe): searching QC again reveals 90 occurrence of this phrase and all of them are confirmed to be Medinan, making it a good target as a distinguishing feature for classification.

(5) Qur'anic initials: There are 30 occurrences of such initial letter opening of chapters of which 3 occurrences are not Meccan, i.e., chapters 2,3, and 13.

(6) Story of Adam and *Iblis*: These are mentioned 5 times in the Qur'an (Q. 2:34, Q. 7:11, Q. 17:61, Q. 18:50, Q. 20:116). I used QC's advanced search that allows a proximity search of input terms which were in our case the words: *Adam* ad *Ibliys*. Only Q. 2:34 is Medinan among this list.

(7) Stories of previous prophets: I used QAC for this search. I selected names of 21 prophets (given in Table 8.x below) which together comprised 581 occurrences. QAC allows search by lemma which neglects the clitics associated with each names and thus gives more accurate results. Usually the results are Meccan but the Medinan chapters no. 2 and 3 have reference to a number of previous prophets.

(8) Linguistic instruments for denunciation and excoriation: To search for this feature I looked for Qur'anic particles that might trigger these senses. For this purpose I again resorted to QAC tagging of particles. This corpus has tags such as: AVR (aversion) usually through the Arabic particle *kalla*, CERT (certainty) using the particle *qad*, SUP (surprise) through the surprise particle *idha*, EXH (exhortation) through the particle *lawla*, and EMPH (emphasis) through the particle *lam*. Using these tags I found 1478 occurrences of these excoriation marks. While most of these marks are indeed characteristics of Meccan chapters, yet there are still quite few instances appearing in Medinan chapters.

(9) Average length of Qur'anic verses: This estimation required a little bit of computation. This involved first downloading the plain Qur'an text from QTP, then creating a separate file for each chapter, then parsing each file and counting the number of words for each verse, and finally finding the average length of the verses of each chapter. Following is a listing of this implementation in the Python programming language.

```
for i in range(1,115):

 ff = "../data/surahs/"+str(i)+".plain"

 s = open(ff,'r')

 line = s.readline()

 noverse = 0.0

 tot = 0.0

 while line!='':

   aya = line.split('|')[2]

   tot += len(aya.split())

   line = s.readline()

   noverse +=1

 print i, ":", tot/noverse
```

Listing 8.1 – Python script to count average length of Qur'anic verses

We found that for the entire Qur'an the average length of verses is around 10 words, and specifically 8 words for Meccan 16 words for Medinan chapters.

(10) Eschatological topics: I choose the following seven words (including clitics) as a marker for this feature: *jahannam* (hellfire), *janna* (paradise), *naar* (fire), *sayeer* (another name for Hellfire), *qiyamah* (day of resurrection), *adhaab* (punishment), and *aakhira* (the world hereafter). I found a total of 822 occurrences of these terms. Like previous cases here again this feature is not exclusively found in Meccan chapters, where some bigger Medinan chapters like no. 2, 3 and 4 has many reference to eschatological terms.

(11) Reference to *Jihad* (religious struggle) through roots words *qtl* and *jhd*: I found 211 such instances in QAC. Although the majority of cases are part of Medinan chapters, Meccan chapters contained some cases as well when: (i) jhd

was used not is the usual sense of Jihad as in *'wa aqsamu billahi jahda aymanihim*' ([Q. 6:109]), or (ii) *qtl* (killing) is referred to the practice of killing their own daughter by the polytheists Arabs as in Q. 6:140, or (iii) referring to killing in the context of stories of previous people like Pharoah killing the male children of Israelites as in Q. 7:127, or brothers of Joseph conspiring to kill him as in Q. 12:9, or (iv) considering the Qur'an as a material for ideological *jihad* as in Q. 25:52.

(12) Medinan chapters detail rulings concerning marriage, breastfeeding and divorce. I choose two root words of: *nkh* (marriage) and *tlq* (divorce). As discussed before considering more words like *zwj* (pair, spous) flags many erroneous terms, and hence I decided to keep only these two terms. This resulted in 37 verses, of which a minority are still noisy like three instance of *intalaqa* (set out) in verses Q. 18:71,74, ad 77.

(13) Reference to the people of the Scripture: I choose five keywords to count this feature: *ahl al-kitab* (people of the scripture), *tawrah* (Torah), *injil* (Gospel), *wahud* (Jews), and *nasara* (Christian). I found 63 instances of which only two instances occurred in Meccan chapters (Q. 29:46 *ahl-alkitab* and Q. 7:157 *tawra wal injil*).

(14) Reference to pillars of Islam: I choose as keyword the four pillars of Islam: *salat* (prayer), *zakat* (alms giving), *siyam* (fasting) and *hajj* (pilgrimage) including any clitics added these words. However, these terms are not exclusively reserved for Medinan chapters though they are the majority. For example, Q. 6:72, 162 mention prayer, Q. 7:156 mentions zakat, Q. 10:87 mentions payer in the context of Moses story when Allah commands him and children of Israel to pray, Q. 18:81 as well as Q. 19: 13 refer to zakat literally to mean purity.

Following table gives a summary of these attributes.

No	Feature	Search for	Category	Exception example
1	Ayat of sajdah (Prostration)	Root sjd	Meccan	Q. 2:144
2	Use of aversion letter kalla	kalla	Meccan	none
3	The phrase: ya ayyiha alladhyna aamanu	ya ayyiha alladhyna	Medinan	none

		aamanu		
4	Qur'anic initials	The tag 'INL' in QAC	Meccan	Suras: 2,3 and 13
5	Story of Adam and Iblis	Proximity search for Adam and Ibliys	Meccan	Q. 2:34
6	Stories of previous prophets	Searched QAC for: 1. <ibora`hiym <isoma`Eiyl 2. yaEoquwb 3. <isoHa`q 4. muwsaY 5. EiysaY 6. daAwud 7. nuwH 8. zakariy~aA 9. yaHoyaY 10. yuwnus 11. ha`ruwn 12. sulayoma`n 13. yuwsuf 14. <iloyaAs 15. yasaE 16. luwT 17. Sa`liH 18. Huwd 19. Adam 20. $uEayob 21. <idoriys	Meccan	Surahs 2, 3
7	Instruments of denunciation and	Following tags from QAC:	Meccan	Q. 2:118

	excoriation	“CERT”, “SUP”, “EXH”, “AVR”, “EMPH”		
8	Average aya length	Meccan avg: 8 words, Medinan average: 16 words per aya		
9	Eschatological topics	Jahannam, jannah, naar, sayeer, qiyamah, adhaab, aakhira	Meccan	Many example in surahs 2,3 and 4
10	Reference to Jihad	Root words qtl and jhd	Medinan	Q 6:109, 140, Q. 7:127
11	Marriage and Divorce	Root words nkh and tlq	Medinan	Q. 18:17,74,77
12	Reference to the people of the Book	Search for: ahl al-kitab, tawrah, injil, wahud, and nasara	Medinan	Q. 29:46, Q. 7:157
13	Reference to the pillars of Islam	Salat, zakat, siyam, and hajj	Medinan	Q. 6:72, Q. 7:156, Q. 10:87

Table 8.2 – Summary of features used to classify Meccan and Medinan Chapters

8.3 Running Experiments using The WEKA Tool

WEKA (Hall et al. 2009) is a tool for data mining that has incorporated various learning schemes and data processing tools. It allows users to quickly try out and compare different machine learning methods on new data sets and visualize results through various graphical means. Thus WEKA enables a convenient data mining platform for non-technical users.

Following is a light walkthrough on the WEKA setup for the problem in our hand, i.e., classification of Meccan and Medinan chapters.

8.3.1 ARFF file

WEKA ARFF file expects the data in a certain format. This file has three parts:

(1) **Relation name**: the first line in the file should be a relation name starting with `@relation` directive, the relation name can be any meaningful name like Meccan-Medinan. So, following is the first part of our file:

```
@relation Meccan-Medinan
```

(2) **Attribute List**: each attribute starts with `@attribute` directive followed by a given name, followed by the type of the attribute. Usually types are either nominal types expecting an enumerated list of outcomes like `{yes, no}` or a number type denoted by the word `'real'`. For example, attribute number 4 in our list is related to Qur'anic initials, and this attribute by nature accepts values either `'yes'` to indicated that this chapter starts with an initial letter, or `'no'` indicating the absence of such initials in that chapter. Similarly, attribute number 5 is also represented as either `yes` or `no` indicating the presence or absence of the story of *Adam* and *Iblis* in that chapter. Other attributes are represented by an actual number of occurrence like attribute number 6 which is represented by an actual count of the names of the 21 prophets in that chapter. Note that the last attribute is a special attribute which tells the expected outcome of the classification of a chapter, which is Meccan or Medinan (for short M and D respectively). Following is the attribute list:

```
@attribute kalla real

@attribute prostration real

@attribute believer real

@attribute initials {YES,NO}

@attribute prophets real

@attribute storyAdamIblis {YES,NO}

@attribute emphasis real

@attribute averageLength real

@attribute paradiseHell real

@attribute jihad real
```

```
@attribute marriage real

@attribute otherReligion real

@attribute pillarsOfIslam real

@attribute place {K,D}
```

(iii) **Date set**: finally the WEKA ARFF file expects a list of data set with each line representing a row of comma delimited list of these attributes. Usually, typical spreadsheets -like Excel- allow conversion of rows into a comma-separated CVS list. Following is the first 10 rows corresponding to the first ten chapters of the Qur'an, note that the list must start with `@data` directive:

```
@data

0,0,0,NO,0, NO,0,4.1,0,0,0,0,0,K

0,12,11,YES,60, YES,62,21.5,55,32,16,8,28,D

0,2,7,YES,27, NO,58,17.51,49,22,0,18,0,D

0,2,9,NO,19, NO,50,21.38,34,28,7,4,12,D

0,1,16,NO,17, NO,53,23.64,20,16,0,18,10,D

0,0,0,NO,23, NO,59,18.52,17,5,0,0,3,K

0,9,0,YES,46, YES,73,16.22,36,3,0,1,2,K

0,1,6,NO,1, NO,15,16.56,10,9,0,0,2,D

0,8,6,NO,8, NO,32,19.42,29,24,0,2,12,D

0,0,0,YES,11, NO,30,16.87,15,0,0,0,1,K
```

This file needs to be saved with `.arff` extension. Once, I have this file complete then I should be able to load it into the WEKA explorer.

8.3.2 WEKA Explorer

Loading our ARFF file in WEKA gives the view in figure 8.1. Note that at the top WEKA has six panels: pre-process, classify, cluster, associate, select attributes, and visualize. I will discuss the functionality of some of these.

8.3.2.1 Pre-process panel

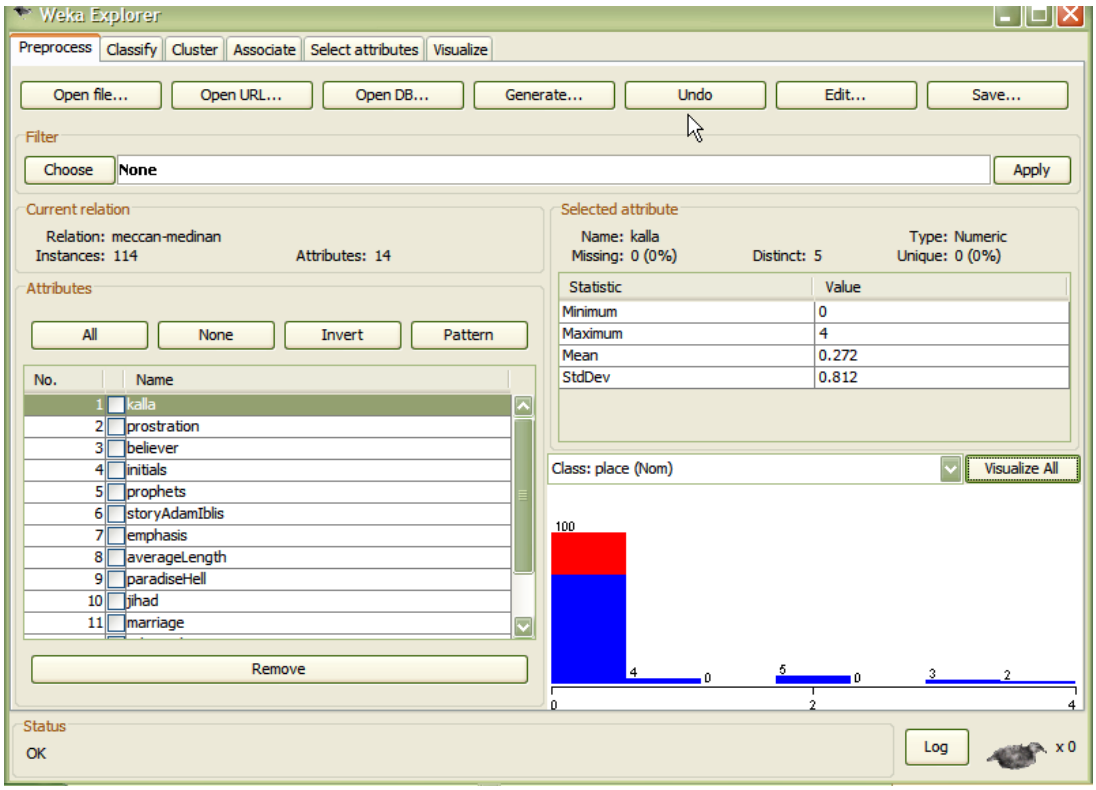


Figure 8.1 - WEKA Explorer pre-processing tab.

WEKA specifies our two classes with two colours, i.e., Meccan as blue and Medinan as red. We can visualize this classification against each of the 13 attributes, or even all of these attribute can be visualized at the same time pressing 'visualize all' button, as shown in figure 8.2.

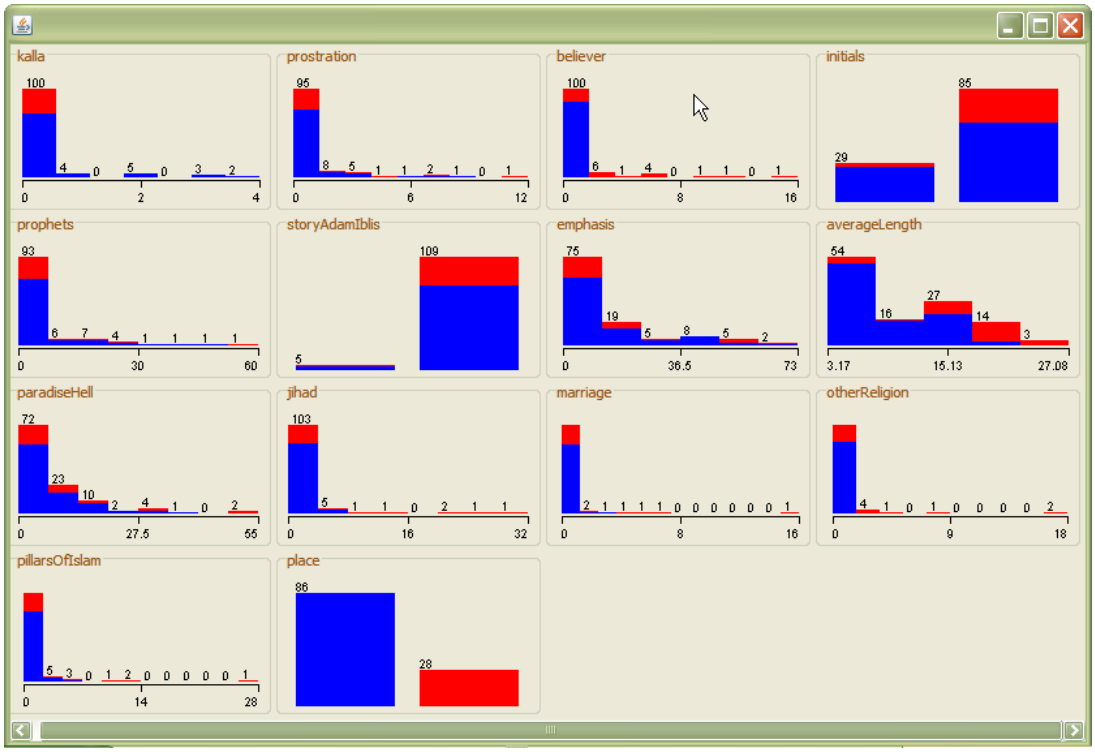


Figure 8.2 – Visualizing all attributes in WEKA

It is very easily evident from this picture the distinctive nature of some attributes: for example, it is easy to verify that attributes: *kalla* has non-zero instance only in Meccan (blue) chapters, and that attributes: *believer* and *otherReligion* has non-zero instances only in Medinan (red) chapters. This panel allows several other functions, like applying various filters (of which we will see an example later), and also editing some values of the attributes.

8.3.2.2 Visualize Panel

This panel allows comparative analysis of the attributes (Figure 8.3). Any two attributes can be selected on a two dimensional x-y graph thus allowing comparison on their correlations.

For example, figure 8.4 gives a comparative picture between the instances of the phrase (*O you who believe*) in the x-axis and those instances that mentions the *four pillars of Islam* in the y-axis. Apart from realizing that all instances of x-axis are Medinan (because all blue colour points have value zero on x-axis), we can also see that there are ten chapters that mention both pillars of Islam and contains this phrase on calling the believer. Further, we can double-click on any point to reveal detailed values. In our case instance number corresponds to chapter number.

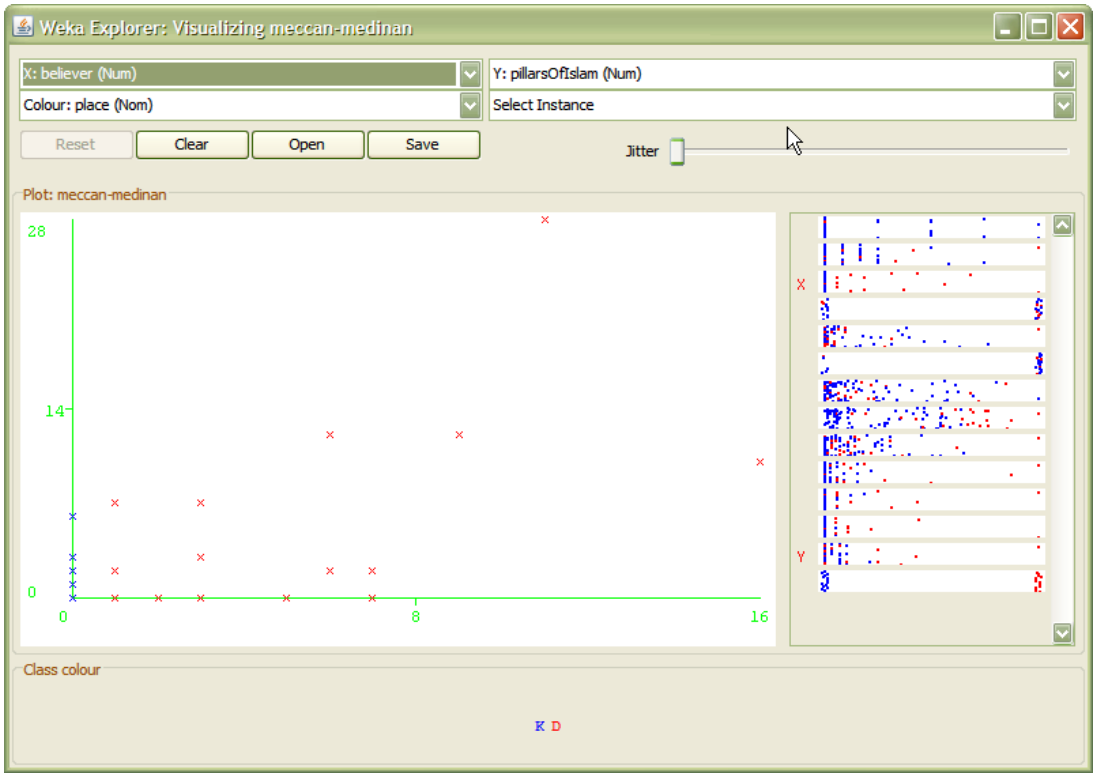


Figure 8.3 - Visualize pane

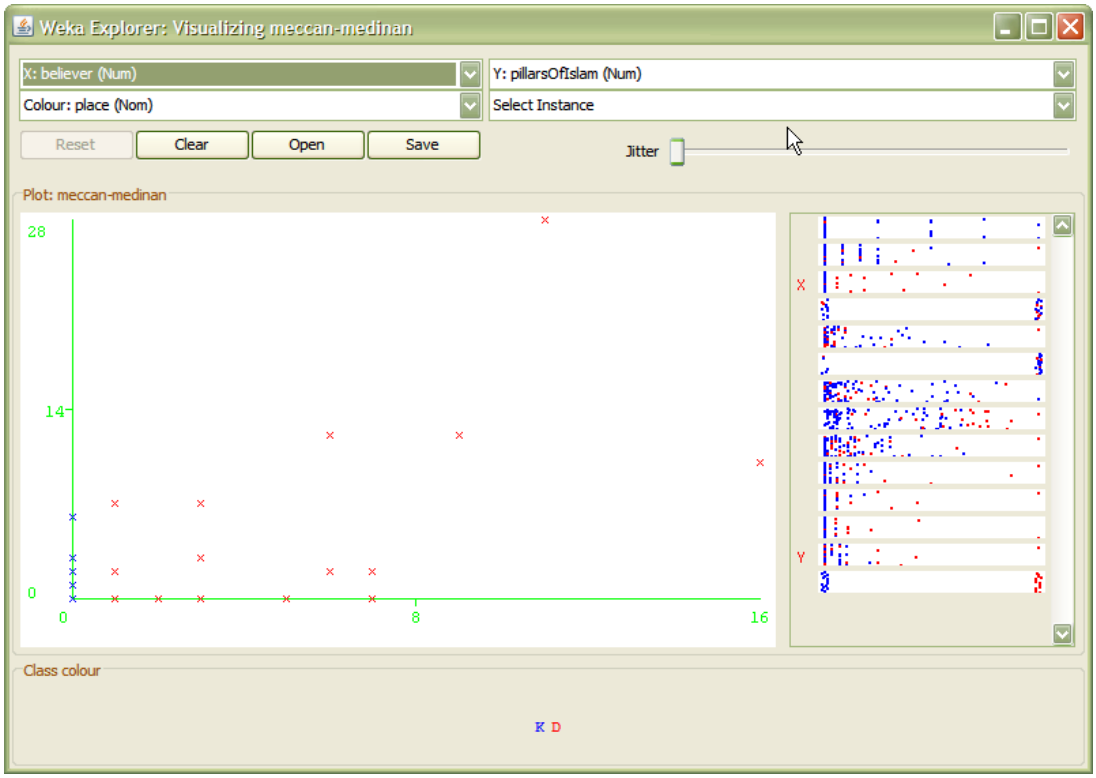


Figure 8.4 – comparing two attributes

8.3.2.3 Classify Panel

Here I will show the main classification task. Two ARFF files are prepared: a training dataset consisting of those 93 chapters that have consensus among scholars based on table 8.1, and another file which contains the testing data consisting of the remaining 21 chapters and we will check to what extend machine was able to classify those debatable chapters.

Figure 8.5 shows the classify pane. First, click on ‘choose’ button. This brings a long list of various Machine Learning classifiers. We expanded the ‘tree’ node and choose the J48 classifier from the list. Next, we select the test option (‘use training set’), then pressing ‘start’ will produce the results on the central pane ‘classifier output’.

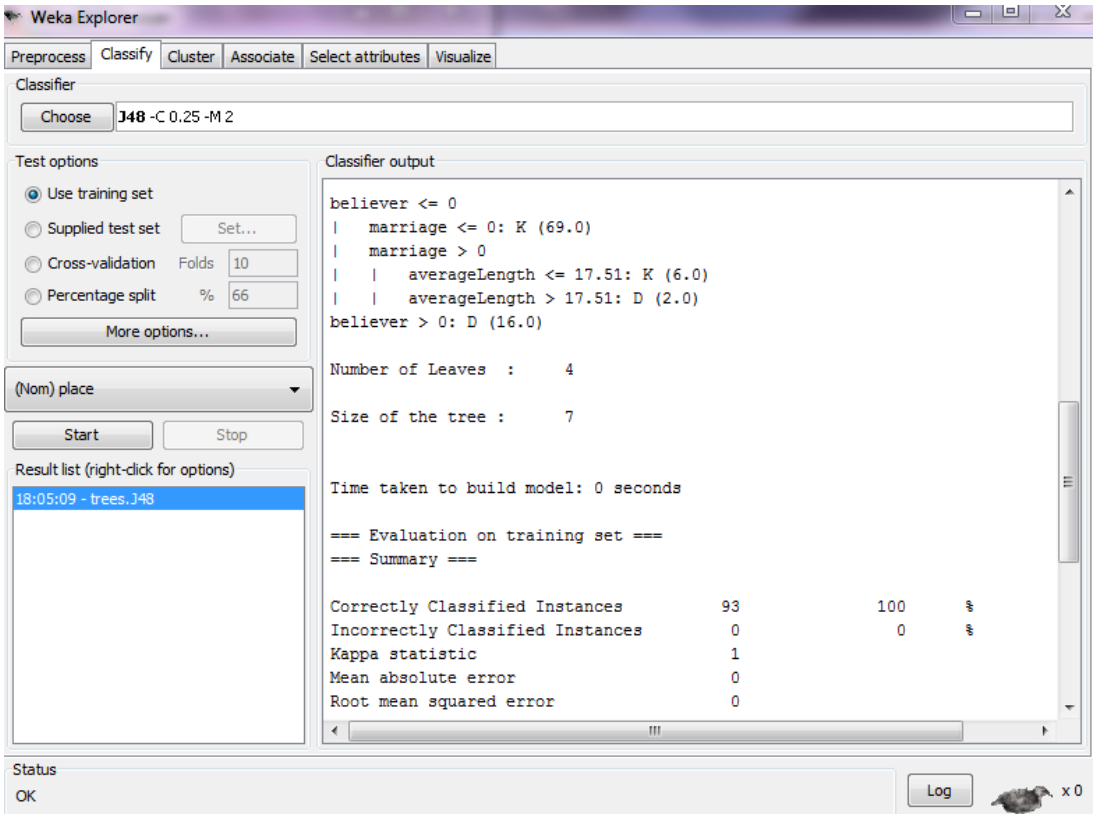


Figure 8.5 - WEKA classification on 93 chapters from the training set

Note that this algorithm could classify all 93 examples correctly and produced a decision tree which can be visualized better by right-clicking on the ‘result list’ and selecting ‘visualize tree’, the result is given in figure 8.6 below.

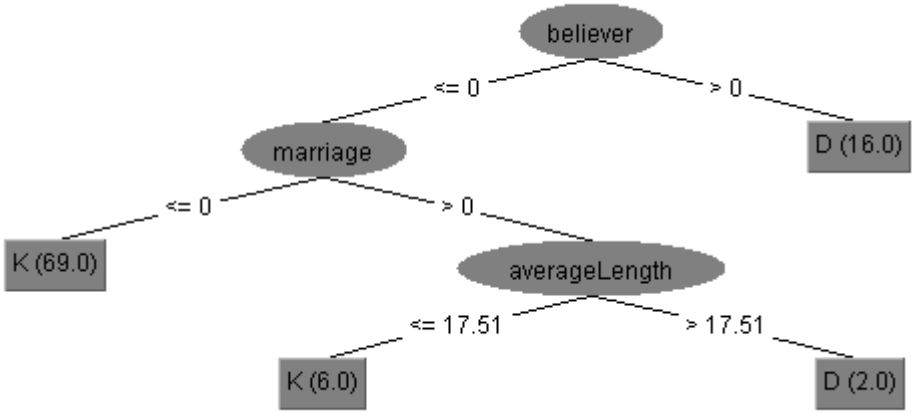


Figure 8.6 - Decision tree using J48 classifier on 93 chapters

The process of generating decision trees can be repeated with different algorithms. For example figure 8.7 below gives the decisions tree based on ‘random tree’ algorithm. Figure 8.8 next page is the decision tree based on ‘Alternating decision tree ADT’ classifier.

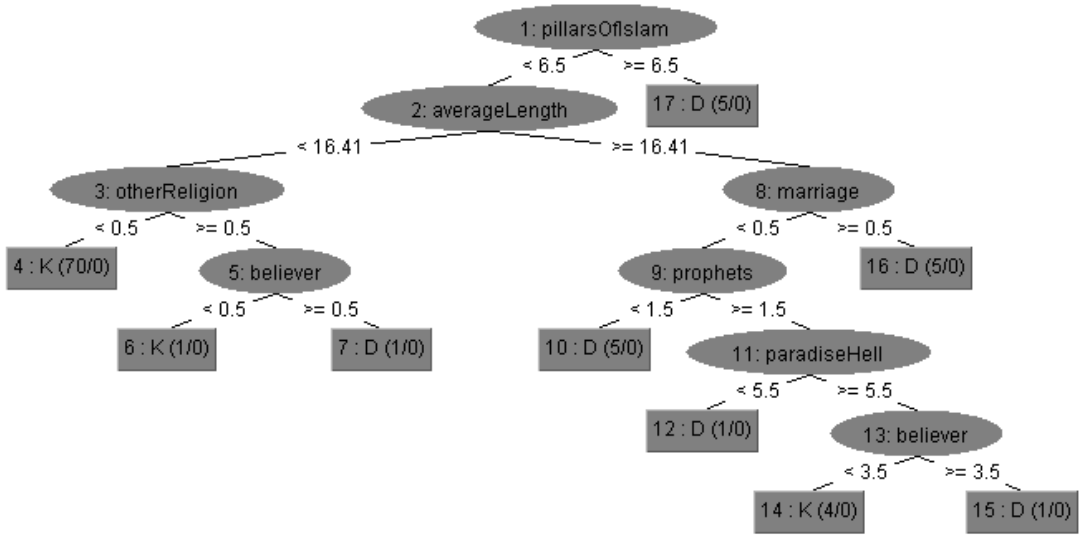


Figure 8.7 - Decision tree produced by ‘random tree’ classifier

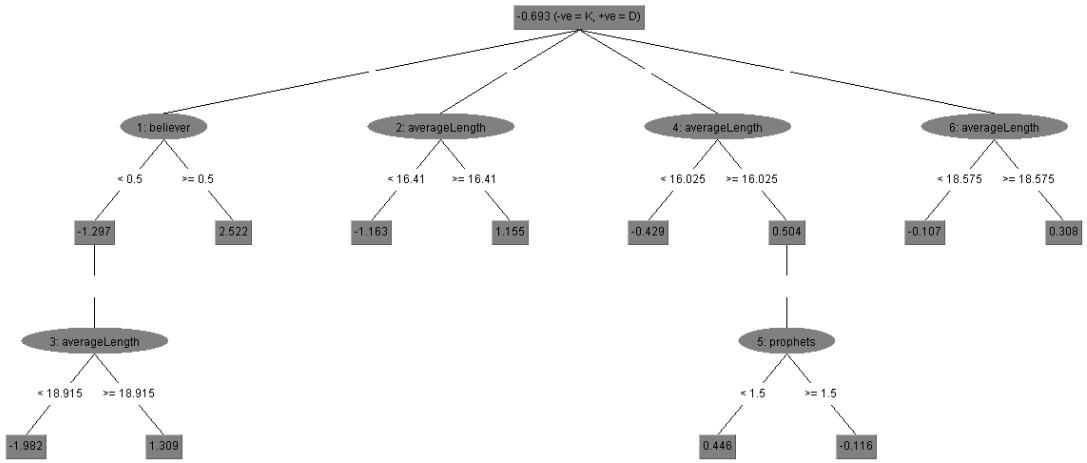


Figure 8.8 - Decision tree based on ADT classifier. Note that negative values indicate Meccan and positive values indicate Medinan

Now, we can check the performance of the algorithm on the 21 debatable chapters. This can be done choosing the 'supplied test set' option from the test options and then choosing the ARFF file for these 21 chapters, and then re-applying the test using 'start' button. To check the prediction of the algorithm for each instance, we checked the 'output predictions' option from the 'more option' list, as shown in the figure 8.9.

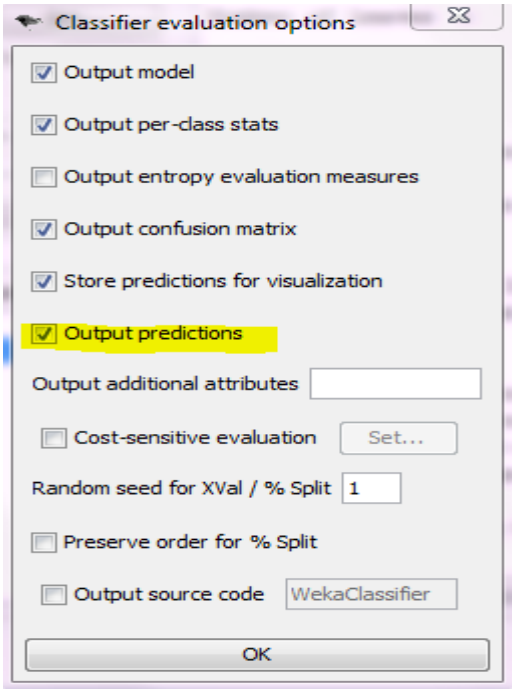


Figure 8.9 - Check 'output predictions' for machine prediction on the 21 test chapters

After this setup, we run the classifier and found out that among the 21 test cases, 6 instances were misclassified. Following are more detailed evaluation statistics as generated by J4.8 classifier.

```
Correctly Classified Instances          15               71.4286 %
Incorrectly Classified Instances         6               28.5714 %
Kappa statistic                          0.4112
Mean absolute error                      0.2857
Root mean squared error                  0.5345
Relative absolute error                 58.8235 %
Root relative squared error             93.6586 %
Total Number of Instances               21
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1          0.6        0.647      1          0.786      0.7       K
                 0.4        0          1          0.4        0.571      0.7       D
Weighted Avg.    0.714      0.314      0.815      0.714      0.684      0.7
=== Confusion Matrix ===
  a  b   <-- classified as
 11  0 |  a = K
  6  4 |  b = D
```

Figure 8.10 shows the Machine Learning prediction based on J48 classifier.

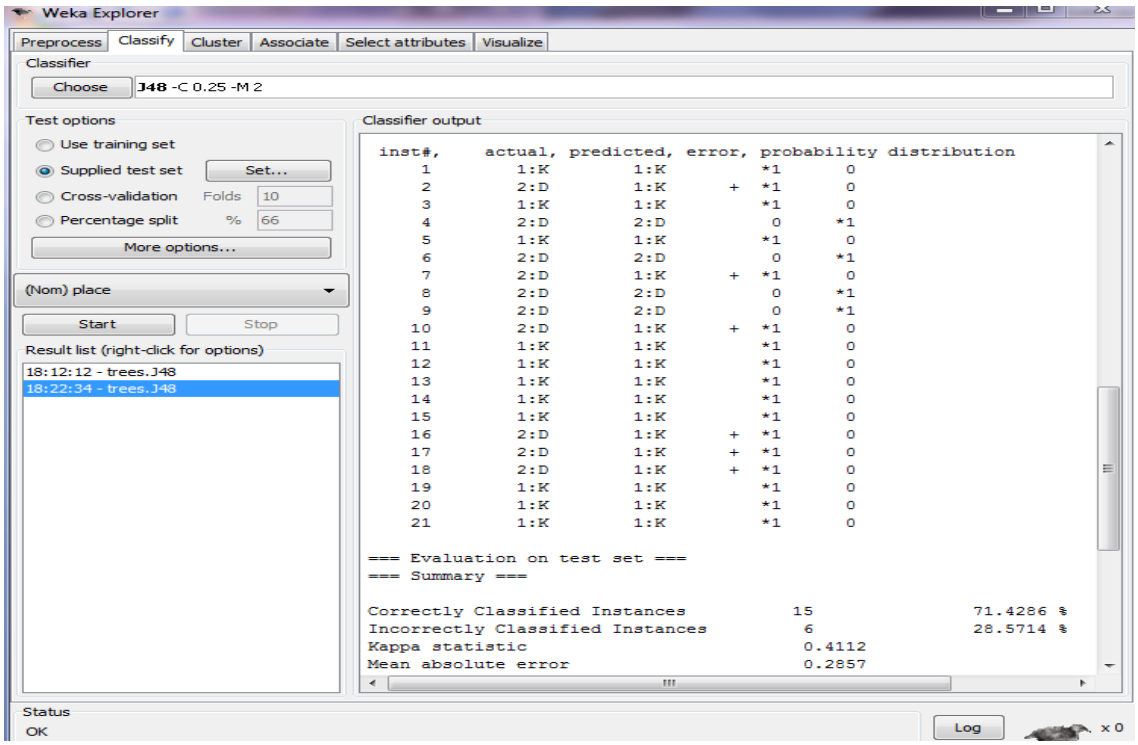


Figure 8.10 - Outcome of the J48 Machine Learning algorithm on the 21 debatable chapters.

Table 8.3 below is a more detailed view of the above outcome; the last column shows the above results, and those predictions that differ from the Medina Qur'an printing complex category are highlighted with grey boxes.

sura no.	kalla	prostrate	believer	initials	prophets	adam-iblis	emphasis	length	eschatology	jihad	marriage	other-religion	pillars	category	prediction
1	0	0	0	NO	0	NO	0	4.1	0	0	0	0	0	K	K
13	0	1	0	YES	0	NO	16	19.84	9	0	0	0	1	D	K
16	0	2	0	NO	4	NO	41	14.41	25	2	0	0	0	K	K
22	0	5	1	NO	6	NO	21	16.4	17	4	0	1	7	D	D
29	0	0	0	YES	13	NO	46	14.17	17	5	0	1	2	K	K
47	0	0	2	NO	0	NO	7	14.26	4	3	0	0	0	D	D
55	0	1	0	NO	0	NO	0	4.51	6	0	0	0	0	D	K
61	0	0	3	NO	3	NO	1	16.14	1	2	0	1	0	D	D
64	0	0	1	NO	1	NO	2	13.44	2	0	0	0	0	D	D
76	0	1	0	NO	0	NO	0	7.84	3	0	0	0	0	D	K
80	2	0	0	NO	0	NO	2	3.17	0	1	0	0	0	K	K
83	4	0	0	NO	0	NO	10	4.69	0	0	0	0	0	K	K
89	2	0	0	NO	0	NO	3	4.63	4	0	0	0	0	K	K
92	0	0	0	NO	0	NO	4	3.38	2	0	0	0	0	K	K
97	0	0	0	NO	0	NO	0	6	0	0	0	0	0	K	K
98	0	0	0	NO	0	NO	2	11.75	2	0	0	2	2	D	K
99	0	0	0	NO	0	NO	0	4.5	0	0	0	0	0	D	K
110	0	0	0	NO	0	NO	0	6.33	0	0	0	0	0	D	K
112	0	0	0	NO	0	NO	0	3.75	0	0	0	0	0	K	K
113	0	0	0	NO	0	NO	0	4.6	0	0	0	0	0	K	K
114	0	0	0	NO	0	NO	0	3.33	0	0	0	0	0	K	K

Table 8.3- Analysis of the outcome of J48 classification of the 21 debatable suras.

Following is a more detailed comment on the outcomes of the Machine Learning algorithm above. We will restrict our discussion to the 6 chapters where our algorithm differed from the classification of the Qur'an printing complex copy.

(1) Chapter no. 13: This chapter contains a number of Meccan chapter characteristics, for example, opening with initial letters '*alim laam raa'*, and it contains a verse of *sajdah* (i.e., prostration), as well as a good number of emphasis and exhortation tools. On the other hand it also contains some characteristics of Medinan chapter, for example it gives reference to prayer as a pillar of Islam, and its verse length is comparable to Medinan chapters.

(2) Chapter no. 55: Although this chapter is considered Medinan by the Qur'an complex, Al-soyuti said about this chapter 'the (the majority) are of the opinion that this chapter is Meccan, and this is the correct opinion'. It is evident that this chapter has a very short verse length (average 4.5 words per verse), and is full of eschatological references.

(3) Chapter no. 76: Again this chapter talks in detail about eschatological issues and its verses are relatively short, and is considered Meccan by many scholars .

(4) Chapter no. 98: A majority of scholars consider this chapter Medinan, and it seems our algorithm only based its judgement on the verse length, and not properly 'learning' from previous examples, that reference to 'people of the book' is a characteristics of Medinan sura.

(5) Chapter no. 99: The scholars verdict on this chapter is that it is Medinan, but based on our feature set suggests it has characteristics of Meccan sūra, especially the short verse length (4.5 words on average).

(6) Chapter no. 110: Again this chapter is classified by our algorithm as Meccan most probably because of its short verse length (6.33 words on average).

8.3.2.4 Clustering

Clustering is a technique in data mining where instances (in our cases Qur'anic Chapters) are to be divided automatically into natural groups. These clustering algorithms attempt to cluster similar instances in one cluster. Thus, it is imperative for these algorithms to incorporate some measure of distance between instances. Various algorithms employ different distance measuring techniques, the common one is called Euclidean distance. Usually these algorithms start to assign a cluster centre called centroid, and then each instance is assigned to the closest centroid. Note that clustering is an unsupervised Machine Learning technique where we only provide our dataset of features without class definitions (for example into Meccan and Medinan), and expect the machine to come up with natural divisions into two or more clusters. These clusters should then be investigated by domain experts (i.e., Qur'anic scholars in our case) in attempt to find interesting correlations between the instances (i.e., Qur'anic chapters in our case). So let us see how WEKA can produce such clusters.

After loading the ARFF file from the 'preprocess' pane, we click on the 'cluster' pane. Then we choose which clustering algorithms we want. We start with '*SimpleKMeans'* algorithm. Users can specify how many clusters they want (the default is 2). Also, users can force the algorithms to ignore some attributes via the 'ignore attribute' button. I decided to ignore the 'place' attribute which specifies each chapter as Meccan or Medinan. In this way we want the algorithms to cluster without knowing which chapter are Meccan or Medinan. Figure 8.11 is the outcome of this algorithm.

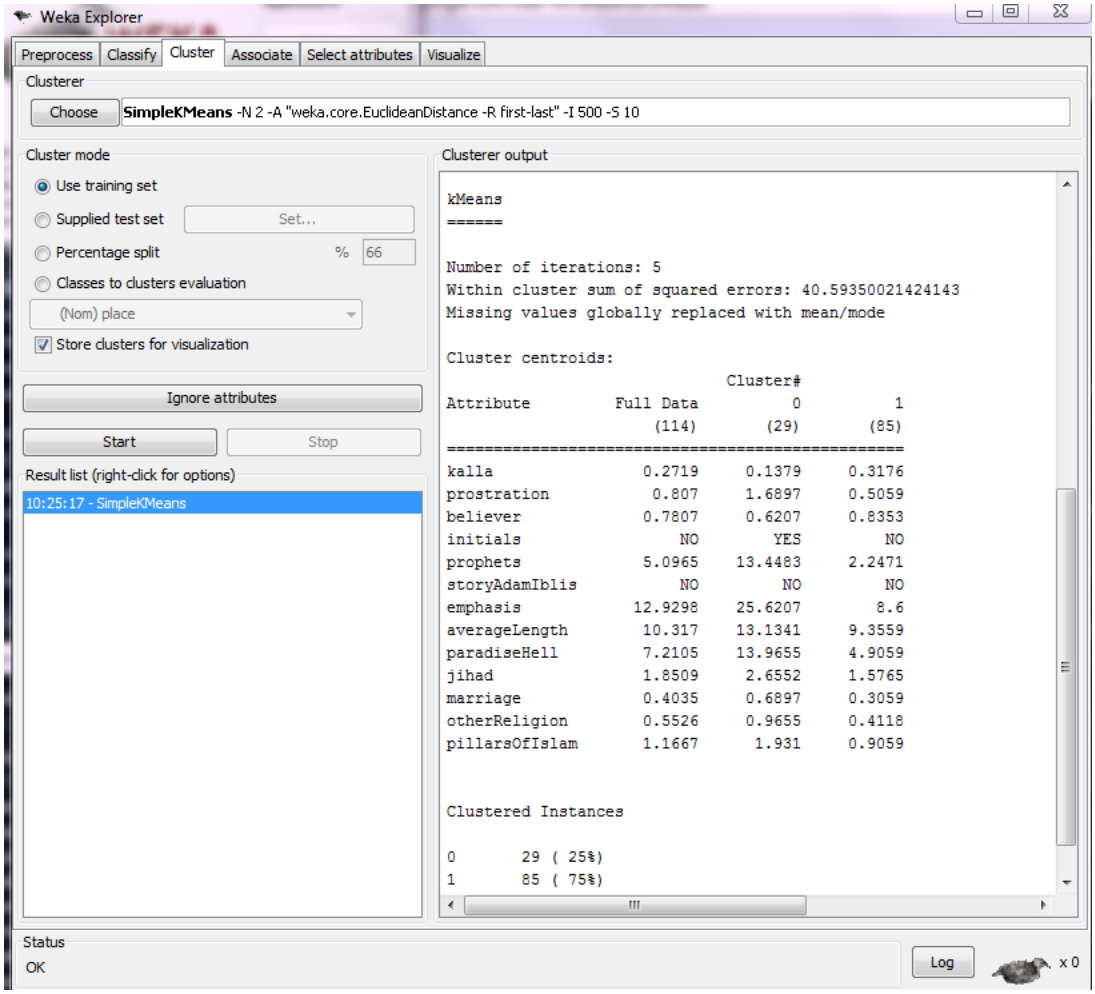


Figure 8.11 - K-Means clustering into 2 clusters

Figure 8.12 below shows visualization graphs of these clusters against the 14 attributes. This can be seen after right-clicking on the results list and choosing '*visualize cluster assignments'*.

From figure 8.12 we can save an ARFF file using the 'save' button. This file can then be loaded into a spreadsheet and manipulated as needed.

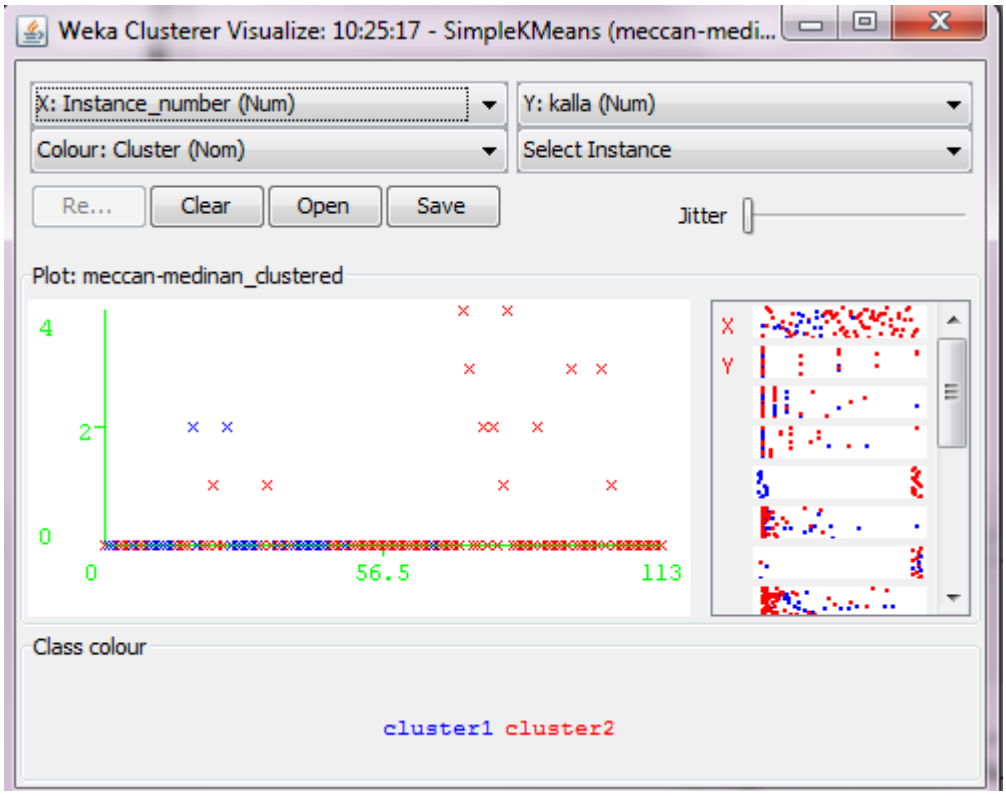


Figure 8.12 - Cluster of the 114 sūra into two clusters against ‘kalla’ attribute

K-Means clustering requires the user to input the desired number of clusters, but that might not be interesting as the user normally would not know beforehand the number of clusters and would be expecting the algorithm to cluster them into optimal groups. For this purpose we used ‘*expectation maximization*’ or EM algorithms from the list, which produced 8 clusters. After saving the ARFF file from visualization, figure 8.13 is a depiction of these clusters. It was interesting to see for example that cluster 1 contained chapters 2 and 7 which has quite a good overlap of content especially in the context of the story of Moses. Also, apart from cluster 7 and 1, Meccan and Medinan classification was preserved.

1		2		3		4		5		6		7		8	
2	D	6	K	15	K	74	K	8	D	1	K	31	K	3	D
7	K	10	K	19	K	83	K	24	D	50	K	35	K	4	D
		11	K	21	K	96	K	33	D	53	K	36	K	5	D
		12	K	23	K					55	D	44	K	9	D
		13	D	26	K					56	K	45	K		
		14	K	30	K					69	K	46	K		
		16	K	32	K					72	K	47	D		
		17	K	34	K					76	D	49	D		
		18	K	37	K					77	K	51	K		
		20	K	43	K					79	K	52	K		
		22	D	54	K					84	K	57	D		
		25	K	68	K					86	K	58	D		
		27	K							88	K	59	D		
		28	K							90	K	60	D		
		29	K							91	K	61	D		
		38	K							92	K	62	D		
		39	K							93	K	63	D		
		40	K							94	K	64	D		
		41	K							95	K	65	D		
		42	K							97	K	66	D		
		48	D							99	D	67	K		
										100	K	70	K		
										101	K	71	K		
										103	K	73	K		
										105	K	75	K		
										106	K	78	K		
										108	K	80	K		
										109	K	81	K		
										110	D	82	K		
										111	K	85	K		
										112	K	87	K		
										113	K	89	K		
										114	K	98	D		
												102	K		
												104	K		
												107	K		

Figure 8.13 - Clustering based on EM algorithm

8.4 Re-running experiments using QurAna dataset

Leveraging on the QurAna dataset I wanted to enhance the counting of some applicable attributes further. For example, I found that a total 1,458 additional mentions of various Prophets are made in the Qur'an through pronouns. Table 8.4 below lists these prophets. Thus our attribute number 6 could be supplemented with more accurate figures.

ID	Concept	count
39	Moses	360
76	Abraham	186
374	Noah	156
579	Joseph	152
59	Jesus	90
467	Lot	64
558	Hud	58
667	Shuaib	51
612	Moses and Aaron	50
602	Solomon	47
304	Zechariah	44
220	Salih	39
82	Jacob	37
686	Aaron	21
146	David	19
604	Ayyoub	16
824	Moses and his servant	13
923	Benjamin	12
77	Abraham and Ishmael	11
610	Ishmael	8
690	David and Solomon	6
613	Elyas	4
789	John	4
401	Noah and Lot	2
445	Noah and Abraham	2
780	Ishmael, Idris and Dhul Kifl	2
804	Idris	2
250	Abraham, Ishmael, Isaac, Jacob and the Descendants	1
611	Abraham and Isaac	1
		1458

Table 8.4 – Count of Prophet names in the Qur'an addressed as pronouns

Next, I re-counted the attribute related to eschatological topics taking into consideration pronoun referents to 12 concepts as listed in table 8.5 below.

ID	Concept	count
12	paradises, gardens	148
24	the bounty of Paradise	2
28	Hellfire	140
168	the day of resurrection	31
186	hellfire 2	2
203	the day of recompense	2
338	the world Hereafter	10
342	people of Paradise and people of Hellfire	5
369	guards of the Hellfire	17
386	people of paradise	83
387	people of Hellfire	51
450	The women of Paradise	7
	total	498

Table 8.5 – Count of eschatological terms in the Qur'an addressed as pronouns

Next, I counted the marriage related terms referred as pronoun antecedent in the Qur'an using QurAna corpus. Table 8.6 below lists the concepts I used resulting in a total of 193 instances.

ID	Concept	count
133	husband	107
105	wife	64
132	a divorced woman	39
119	husband and wife	22
117	divorced women	9
	total	241

Table 8.6 - Count of 'marriage' related terms in the Qur'an addressed as pronouns

Next, I counted the reference to the 'people of other scriptures' and in this regard I considered the list of concepts as detailed in table 8.7 below. Thus, a total of 788 new counting items are now added.

ID	Concept	count
52	Jews	288
54	People of the Book	202
51	the Torah	288
1042	the Gospel	3
30	Torah and Gospel	7
310	Christians	64
68	Jews and Christians	68
	total	788

Table 8.7 - Count of the concepts related to ‘People of the Book’ in the Qur’an addressed as pronouns

Finally, I counted the concepts related to ‘the pillars of Islam’ and accumulated the following list of concepts detailed in table 8.8 totalling to 34 concepts. We notice from this table low number of counts, as these pillars are acts of worship, and as such it is not a likely to refer to an act of worship through a pronoun.

ID	Concept	count
763	pilgrims	13
948	those who do not give alms	10
289	those who believe, does good, establish prayer and give zaka	6
31	the Prayers	4
259	fasting	1
	total	34

Table 8.8 - Count of the concepts related to ‘pillars of Islam’ in the Qur’an addressed as pronouns

After accumulating these enhanced attributes from QurAna, we re-run the experiments using the WEKA tool as described in the previous section 8.3. Results show improvements achieved through using QurAna. Figure 8.10 shows that among the 21 test instances the J4.8 classifier misclassified 6 instances making accuracy of 71.2% with a combined F-Measure of 0.684. Leveraging on QurAna J4.8 classifier was able to classify 16 instances correctly reducing the misclassified suras to only 5 with accuracy level of 76.2% and average combined F-Measure of 0.744. Following is a listing of detailed evaluation output from WEKA.

```
Correctly Classified Instances          16               76.1905 %
Incorrectly Classified Instances         5               23.8095 %
Kappa statistic                          0.5116
Mean absolute error                      0.2381
Root mean squared error                  0.488
Relative absolute error                 49.0196 %
Root relative squared error             85.4982 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.5        0.688     1         0.815      0.75     K
                 0.5       0          1         0.5       0.667      0.75     D
Weighted Avg.    0.762     0.262      0.836     0.762     0.744      0.75

=== Confusion Matrix ===

  a  b   <-- classified as
 11  0 |  a = K
  5  5 |  b = D
```

Figure 8.14 below is a screen capture of the classification pane where prediction of the learning algorithm on the test set is given for each instance.

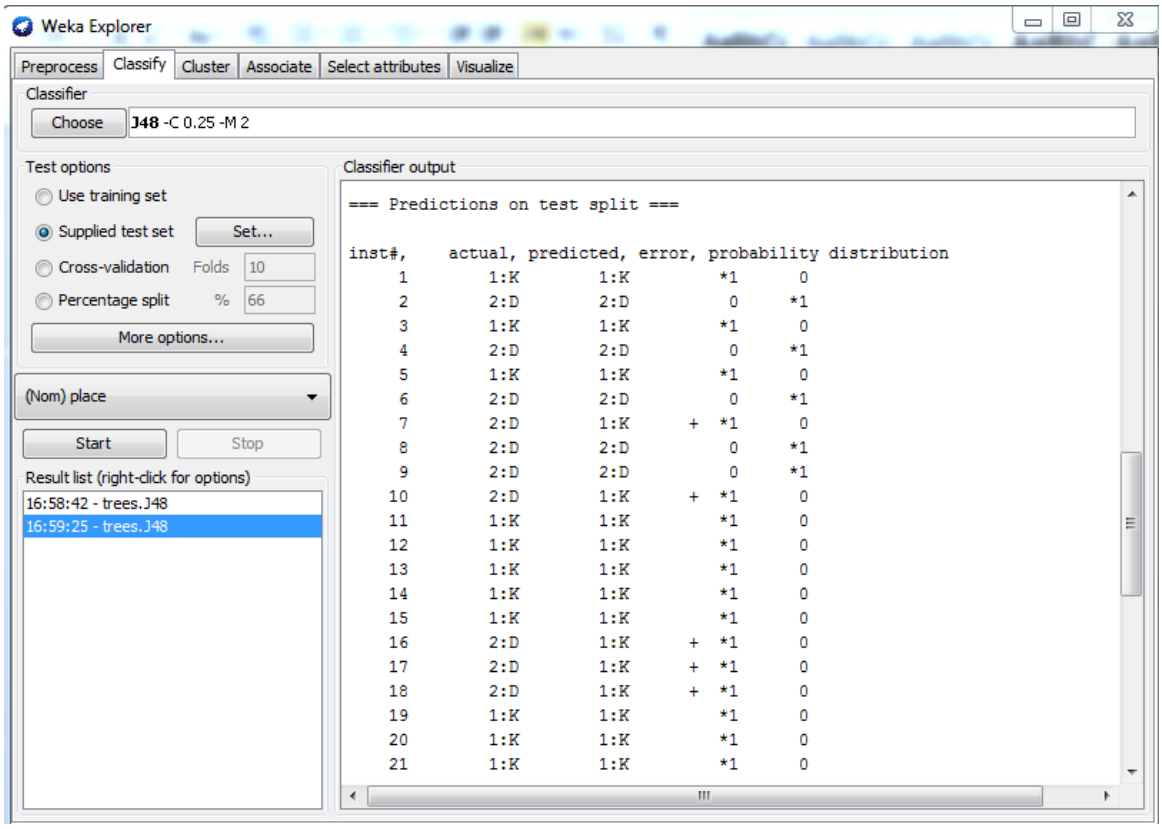


Figure 8.14 - Outcome of the J48 Machine Learning algorithm on the 21 debatable chapters after incorporating QurAna corpus

Table 8.9 is a revisit of table 8.3 where prediction of the classifier is given both before and after incorporating QurAna.

sura no.	kalla	prostrate	believer	initials	prophets	adam-iblis	emphasis	length	eschatology	jihad	marriage	other-religion	pillars	category	prediction-1	prediction-2
1	0	0	0	NO	0	NO	0	4.1	0	0	0	0	0	K	K	K
13	0	1	0	YES	0	NO	16	19.8	18	0	0	3	1	D	K	D
16	0	2	0	NO	8	NO	41	14.4	29	2	0	9	0	K	K	K
22	0	5	1	NO	7	NO	21	16.4	35	4	0	2	20	D	D	D
29	0	0	0	YES	42	NO	46	14.2	29	5	0	6	2	K	K	K
47	0	0	2	NO	0	NO	7	14.3	13	3	0	2	0	D	D	D
55	0	1	0	NO	0	NO	0	4.51	14	0	0	0	0	D	K	K
61	0	0	3	NO	11	NO	1	16.1	2	2	0	3	0	D	D	D
64	0	0	1	NO	1	NO	2	13.4	5	0	0	0	0	D	D	D
76	0	1	0	NO	0	NO	0	7.84	10	0	0	0	0	D	K	K
80	2	0	0	NO	0	NO	2	3.17	0	1	0	0	0	K	K	K
83	4	0	0	NO	0	NO	10	4.69	2	0	0	0	0	K	K	K
89	2	0	0	NO	0	NO	3	4.63	4	0	0	0	0	K	K	K
92	0	0	0	NO	0	NO	4	3.38	4	0	0	0	0	K	K	K
97	0	0	0	NO	0	NO	0	6	0	0	0	0	0	K	K	K
98	0	0	0	NO	0	NO	2	11.8	5	0	0	5	2	D	K	K
99	0	0	0	NO	0	NO	0	4.5	0	0	0	0	0	D	K	K
110	0	0	0	NO	0	NO	0	6.33	0	0	0	0	0	D	K	K
112	0	0	0	NO	0	NO	0	3.75	0	0	0	0	0	K	K	K
113	0	0	0	NO	0	NO	0	4.6	0	0	0	0	0	K	K	K
114	0	0	0	NO	0	NO	0	3.33	0	0	0	0	0	K	K	K

Table 8.9 – Attribute counts of 21 test cases for Meccan-Medinan classification. Prediction-1 shows classifier result before QurAna counts, and Prediction-2 is after incorporating QurAna counts. Shaded cells show the misclassification instances. Shaded columns are the attributes where QurAna was incorporated.

We notice from the table that the sura no. 13 was the only sura which was not correctly classified before is now correctly classified after incorporating QurAna. When investigating individual counts of attributes, we notice that two attributes were influenced by QurAna counts: first, 'eschatological' terms doubled from 9 to 18 and second, 'reference to other religion' showed 3 positive cases using QurAna after being none before. We knew from scholarly observation that while the first attribute favours Meccan category the second favours Medinan, and the classifier in this case gave more weight to the 'other-religion' attribute and correctly classified this sura.

Figure 8.15 shows the decision tree generated by WEKA. Comparing this with an earlier tree (figure 8.6) we notice that both start branching on the attribute 'believer', however, the second branch in QurAna case is on 'average Length' of verses, whereas the earlier case is was on terms related to 'marriage'.

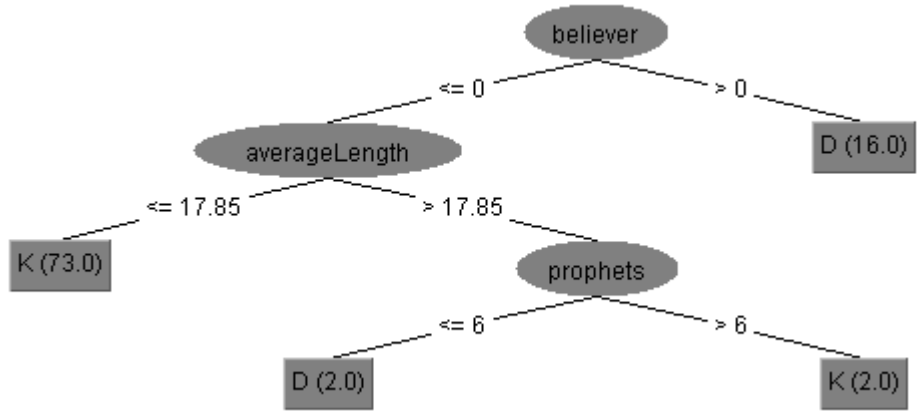


Figure 8.15 – Decision tree using J48 classifier after incorporating QurAna

8.5 Experiments for calculating verse distance using QurAna and QurSim

The vector space model is widely used in information retrieval where the query terms and each document are represented as vectors and the distance between query and document is measured by comparing the cosine of the angle between them. I followed the same methodology and considered each verse of the Qur'an as a separate document.

Each verse was then modelled as a term vector taking roots of the Qur'an as the terms. The Qur'an has 1,226 unique roots, from these I have kept roots repeated over 2 times, and removed the first 3 most frequent roots. Thus, our vector for each verse contains 758 roots as term indices. Next, in order to give a weight for each term, I used term frequency – inverse document frequency (tf-idf) metric, using the following formula adapted from (Sebastiani 2002).

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{\#_{Tr}(t_k)}$$

Where *#(tk,dj)* denotes the number of times the root t_k occurs in the verse d_j, and # $T_r(t_k)$ denotes the verse frequency of root t_k, that is, the number of verses in the Qur'an T_r in which the root t_k occurs.

In order for the weights to fall in [0,1] interval and for the verses to be represented by vectors of equal length, the weights (w_{kj}) resulting from *tfidf* were normalized according to the following formula for cosine normalization:

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|}(tfidf(t_s, d_j))^2}}$$

To find the distance (or measure of similarity) between two vectors, cosine angle is measured using the formula below, where A, B denotes two verses' vectors:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Similarity values fall between [0,1], where 0 indicates no similarity, and 1 indicates identical matching. Using the above setup, I have evaluated the QurSim dataset of 7,679 related verse pairs and found out that only 428 pairs (6%) produced a similarity value above 0.5. This finding confirms the assumption that automatic computation of verse relatedness requires integration with domain specific knowledge sources and relying only on lexical matching produces poor results.

Given these results I considered next how to enrich a verse vector with concepts from our ontology. Instead of constructing root vectors for a verse from only that verse's root, I augmented this verse's roots with roots of all other verses that share common antecedent. For example consider verse 27:26 below:

بَلِ ادَّارَكَ عِلْمُهُمْ فِي الْآخِرَةِ ۚ بَلْ هُمْ فِي شَكٍّ مِّنْهَا ۖ بَلْ هُم مِّنْهَا عَمُونَ

Nay, / is arrested / their knowledge / of / the Hereafter? / Nay / they / (are) in / doubt / about it. / Nay, / they / about it / (are) blind. /

Nay, but doth their knowledge reach to the Hereafter? Nay, for they are in doubt concerning it. Nay, for they cannot see it.

Qur'anic Verse – 27:66

This verse contains 3 concepts marked by pronoun referents: 'the polytheists', 'those who deny resurrection' and 'the world Hereafter'. Therefore, I have augmented the term vector of the verse 27:66 with the terms from all other verses that have any of these three concepts.

The similarity measurement experiment described above was repeated using these improved vectors, and the same dataset was used. While in the early experiment, only 428 pairs showed a similarity distance over 0.5, augmenting verses with their concepts showed 869 pairs from the total of 7,679 pairs in our dataset, i.e., over 50% improvement.

8.6 Summary

This chapter described some machine learning experiments performed on the Qur'an. The machine learning problem chosen in this work is to classify Qur'anic chapters according to their chronological order into Meccan and Medinan chapters. Significance of this problem for Qur'anic studies was described in detail. Rich sets of linguistic and domain specific features were incorporated in the learning process extracting from scholarly works. The WEKA tool was used to run experiments. The classifier was tested on a set of 21 Qur'anic chapters that are disputed among scholars on their classification. Although running test cases on non-disputed cases would give better results, in this experiment I chose disputed cases to demonstrate the usefulness of machine learning experiments

to Qur’anic scholars. The results showed an accuracy level of 71.4%. When the experiments were run again incorporating QurAna corpus in counting the features, the accuracy level increased to 76.2%.

The chapter also described machine learning experiments done to predict the similarity of two verses using as training set vector space model and lexical similarity measures of Qur’anic verses. This experiment was repeated after incorporating QurAna concepts to the vector measures and observed 50% improvement.

Chapter 9
Conclusion and Future Work

9.1 Overview

9.1.1 Overall findings

Muslims believe that the Qur'an is a sole authoritative source for knowledge, wisdom, guidance and legislations for mankind. It has been a great challenge for the computer scientists to represent the knowledge embodied within this text and develop intelligent systems that can extract knowledge from the Qur'an. In this thesis I developed a number of useful resources that were exploited for text mining the Qur'an. I have developed two useful language resources based on the Qur'an: QurAna and QurSim. QurAna contains tagging of the 24,000 Qur'anic pronouns with their antecedents. QurSim is a dataset of around 8,000 related verse pairs compiled from scholarly sources. I have introduced the novel idea of maintaining a register of ontological concepts out of pronoun referents as I progressed with annotating task.

These resources were then used for a number of custom made text mining applications that were placed online for public use. Among these applications: lemma concordance, collocation, POS search of the Qur'an, verse similarity measures, concept clouds of a given verse, pronominal anaphora and Qur'anic chapter similarity.

Moreover, supervised machine learning techniques were used for automatic classification of Qur'anic chapters benefiting from linguistic features dictated by Qur'anic scholars. The accuracy of the classifier was improved after incorporating QurAna corpus data.

9.1.2 Chapter Summaries

Chapter 1

This chapter provided an introduction to this research which focuses on building language resources that would eventually enter into beneficial text mining applications. The main contribution of this thesis has been the development of three novel resources, namely, a corpus of Qur'anic pronoun tagging, a domain concept ontology of the Qur'an emanating from pronoun referents and a dataset of related verses in the Qur'an compiled from scholarly sources. These resources

were used in a number of text mining applications and machine learning experiments.

This chapter explored a number of possible ways this research could add value to the NLP community, particularly: Arabic NLP, computational text similarity analysis, computational anaphora resolution, stylometrics, and translation studies.

Chapter 2

This chapter discussed the rationale behind choosing the Qur'an and introduced a number of topics related to the domain investigated by this thesis: i.e., the Qur'an. First, the Qur'anic language – Classical Arabic – is compared with Modern Standard Arabic. Next, a number of unique linguistic characteristics of the Qur'an was discussed for example, the inimitability of the Qur'an, scattered information on the same subject, verb preposition binding, metaphor and figurative use.

Also, this chapter introduced linguistic background of the linguistic subjects investigated in this thesis, namely the pronominal anaphora system in Arabic, and the text similarity and relatedness in the Qur'an.

As I resorted to both Qur'anic exegesis and translations in our research, this chapter also introduced the significance of these two topics.

Chapter 3

This chapter provided a literature review of some computational works carried on the Qur'an. For example, (Thabet 2005) and (Moisl 2009) worked on forming clusters of Qur'anic chapters based on statistical distribution of some keywords. (Sadeghi 2011) studied the chronology of Qur'anic chapters by employing stylometric measures.

Major anaphora resolution systems were also introduced in this chapter. Finally, computational text similarity and relatedness was covered by reviewing major evaluation corpora and relatedness measures.

Chapter 4

This chapter introduced a number of existing annotations of the Qur'anic texts. Some of the reviewed annotations has exhaustively covered the entire Qur'an – for example, our QurAna corpus or the corpus of pause marks of the Qur'an, and some are partial attempts, for example, syntactic annotation Treebank available at the Qur'anic Arabic Corpus website.

Haifa corpus and Qur'anic Arabic Corpus exhaustively analysed the morphology of every word of the Qur'an. However, QAC went through a rigorous verification process and adapted a collaborative platform for further evaluation.

QurAna picked up pronoun tags from QAC and annotated their referents. Creation of a concept ontology while tagging pronoun referents is a novel experiment I carried through this research. This approach proved helpful as the Qur'an is loaded with many pronouns.

This chapter included a brief description of some partial annotation attempts to capture semantics of the Qur'an, like the Qur'anic semantic frames and Qur'anic prepositional verbs.

Finally, this chapter benchmarked the Qur'an against typical characteristics of a corpus: sampling and representativeness, finite size, machine-readable form and a standard reference.

Chapter 5

This chapter discussed in detail the QurAna corpus which captures the annotation of over 24,000 Qur'anic pronouns with their antecedents and maintains in parallel an ontology of Qur'anic concepts from these antecedents. The annotation schema employed for building QurAna is comparable to other schema designed for similar tasks like the UCREL schema or MUC-7 SGML schema.

The chapter described how QAC was integrated with the annotation process and how available scholarly comments on the Qur'an were helpful in resolving ambiguous cases. This corpus along with concept ontology was incorporated into online applications where users can query this corpus.

Finally, this chapter discussed some challenges faced during the annotation process and future improvements that could enhance QurAna.

Chapter 6

This chapter presented QurSim: a large corpus created from the original Qur'anic text, where semantically similar or related verses are linked together. This dataset can be used for evaluation of paraphrase analysis and machine translation tasks. QurSim is characterised by: (1) superior quality of relatedness assignment; as QurSim has incorporated relations marked by well-known domain experts, this dataset could thus be considered a gold standard corpus for various evaluation tasks, (2) the size of QurSim; over 7,600 pairs of related verses are collected from scholarly sources with several levels of degree of relatedness. This dataset could be extended to over 13,500 pairs of related verses observing the commutative property of strongly related pairs.

This dataset was incorporated into online query pages where users can visualize for a given verse a network of all directly and indirectly related verses. Empirical experiments showed that only 33% of related pairs shared root words, emphasising the need to go beyond common lexical matching methods, and incorporate -in addition- semantic, domain knowledge, and other corpus-based approaches.

This chapter concluded with describing some challenges faced during the compilation process and suggested some ways to improve QurSim in future.

Chapter 7

This chapter described some text mining applications made online for querying the Qur'an. These application were made possible because of the development of various language resources and annotation layers of the Qur'an including QAC, QurAna and QurSim.

Among the text mining applications described in this chapter are: Qur'anic concordancer, QurAna searching application, lexical similarity measures of Qur'anic verses, QurSim query and visualization, Qur'anic concept clouds, semantic relations between Qur'anic chapters, n-gram search of the Qur'an, Qur'anic word co-occurrence, part-of-speech visualization of the Qur'an, Qur'anic chapter word cloud, and concept ontology from pronoun referents.

Chapter 8

This chapter described some machine learning experiments performed on the Qur'an. The machine learning problem chosen in this work is to classify Qur'anic

chapters according to their chronological order into Meccan and Medinan chapters. The significance of this problem for Qur'anic studies was described in detail. Rich sets of linguistic and domain specific features were incorporated in the learning process extracting from scholarly works. The WEKA tool was used to run experiments. The classifier was tested on a set of 21 Qur'anic chapters that are disputed among scholars on their classification, and results showed 71.4% accuracy level. When the experiments was run again incorporating the QurAna corpus in counting the features accuracy level increased to 76.2%.

The chapter also described machine learning experiments done to predict similarity of two verses using as training set vector space model and lexical similarity measures of Qur'anic verses. This experiment was repeated after incorporating the QurAna concepts to the vector measures and observed 50% improvement.

9.2 Aims and Objectives

The aim of this thesis is to apply tools and techniques in text mining to the field of the Qur'anic studies. The objective of this thesis was then to develop some language resources and text mining tools that would provide a case study for further research in this field.

Throughout the course of this thesis, I showed how this aim and objective was fulfilled. Two novel resources were created and made available for research purpose: the QurAna corpus with Qur'anic pronouns annotated with referents and the QurSim dataset where nearly 8,000 pairs of related Qur'anic verses were compiled.

A number of text mining tools and applications were designed and deployed through online query pages leveraging on these developed datasets. The developed resources and applications were made public to encourage research in this field. Since then I have received a great deal of feedback and comments, and a lot of interest has been expressed from the research community.

We have successfully participated in a workshop arranged by the British Computer Society for the Grand Challenges for Computing Research in 2010 under the title "Understanding the Qur'an: A new Grand Challenge for Computer Science and Artificial Intelligence" (Atwell et al., 2010). Following is an excerpt.

> *Access to the Qur'an has traditionally been through the text: many Muslims learn to memorise and recite the verbatim data-set. Access to the*

> *underlying knowledge, wisdom and law requires interpretation and inference; much knowledge is encoded via subtle use of words, grammar, allusions, links and cross-references. For over a thousand years, scholars have sought to extract knowledge and laws from the text, and have built up a much larger Tafsir or corpus of analyses, interpretations and inference chains. Computer Science and Artificial Intelligence presents the opportunity to re-analyse the text data, extract and capture the underlying knowledge in a Knowledge Representation and Reasoning formalism, and enable automated, objective inference and querying.*

This research thesis via the developed resources and applications has pioneered in contributing towards fulfilling this grand challenge for computing and artificial intelligence.

9.3 Future works

This research laid a foundation for many future tasks. A major portion of my PhD time and effort went into the task of annotating the 24,000 pronouns of the Qur'an, and compiling the dataset of nearly 8,000 pairs of related verses. After having these resources at my disposal a number of text mining applications and machine learning experiments were conducted. I left out a number of possible extensions, applications and machine learning experiments as future improvements. The following subsections provide an outline.

9.3.1 Improvements on QurAna and QurSim

As both QurAna and QurSim were developed as the first version, in future both of these dataset can be improved further in a number of ways. I have outlined a number of potential usage of QurAna and QurSim in section 1.3. Detailed discussion on challenges and future enhancements were made in section 5.6 for the QurAna and section 6.5 for the QurSim dataset. Here I discuss more generic future enhancements in relation to these resources.

Validation: These datasets were created resorting to scholarly works and hence one would assume they were accurate and verified. However, they need to be subjected to further manual validation by Qur'anic scholars before incorporating them into wider applications.

Extension: in order to allow for wider research use in the field of Arabic NLP, I would suggest extending these datasets to incorporate texts both form a similar register (e.g., incorporating traditions of the Prophet Muhamamd i.e., Hadith) and corpora from Modern Standard Arabic (MSA).

Further layers of annotation: as of now, morphological (i.e., QAC) and pronoun (i.e., QurAna) annotations are available for the entire Qur'an. Few more layers of annotation are available partially (e.g., syntactic annotation using dependency treebanks), and few annotations are not available as machine readable format (e.g., annotation of Tajweed rules ref. 4.4.2). completing these layers of annotation would enable more through text mining application and knowledge extraction from the Qur'an.

Text Mining Applications: a number of text mining applications were made as part of this thesis (e.g., see chapter 7). In future more applications can be made leveraging on QurAna, QurSim and other developed layers of annotation. For example, I discussed some computational stylistics features of the Qur'an. With the availability of QurAna dataset, we could correlate stylistic features pertaining to the usage of pronouns in the Qur'an. Also, as we have now QurAna dataset, machine learning experiments can be done using this dataset for automatic anaphora resolution system. Similarly, with the QurSim dataset, experiments for automatic detection of related text can be performed.

Dataset for Shared Tasks in Computational Semantics: QurSim can be used as an evaluation dataset for a number of interesting challenges in the field of computational semantics. QurSim enjoys the use of the relatively large dataset and can easily be deployed to multiple languages. In fact, I contacted the organizers of *SEM Shared Task 2013 (SEM 2013) for the feasibility of proposing a task based on QurSim and I received positive feedback and was advised to submit a formal proposal.

The next section (9.3.2) introduces a few potential applications that are of interest to Qur'anic scholars. .

9.3.2 Potential Machine Learning applications to Qur'anic Studies

We have shown in Chapter 8 an example of useful machine learning experiment in the context of Qur'anic studies. The general approach towards employing data mining techniques is: to first define an interesting problem where automation and Machine Learning can prove helpful. These problems are usually characterized by the availability of a dataset with distinguishing quantifiable features which has the potential of embedded hidden patterns and correlations in the dataset. Machine Learning techniques come to reveal such patterns and associations through decision trees, automatics clustering and associations. The outcome of

such algorithms is to be interpreted by a Qur'anic scholar in order to evaluate its usefulness.

Following are some suggested potential applications of Machine Learning and data mining techniques to the field of Qur'anic studies. I have attempted to relate these applications to previous research published in the Journal of Qur'anic Studies to advocate the relevance of our computational methodology for Qur'anic studies. These suggested applications enjoy the same characteristics of our detailed Meccan and Medinan task explained in this paper: first the problem at hand needs to be reduced into countable number of features, then a dataset of examples need to be created for this problem and fed into Machine Learning algorithms which can learn from this training set, then depending on the nature of the application a testing set can be entered to predict the machine outcome, or a set of clustering and association rules can be investigated by Qur'anic scholars.

9.3.2.1 Discovering patterns from Prophet's companion's exegesis

(Geissinger 2004) examined in her paper some traditions from the most popular Hadith source *Al-Bukhari* that show exegetical comments by Aisha (i.e., Prophet Muhammad's wife) on Qur'anic verses related to theology as well as the ritual of *Hajj* (pilgrimage to Makkah). Machine Learning algorithms can be employed to investigate classical exegesis books like *Ibn Katheer* or *Tabari* and collect all *hadiths* that are attributed to the prophet's companions. In this way we end up with a large dataset in the form of a matrix of companions against the Qur'anic verses on which these companions commented. This matrix can then be analysed by Machine Learning algorithms to find clusters and interesting associations and patterns that might highlight certain thematic exegetical preference pertaining to certain companions. For example, a row in the matrix would be associated with *Aisha* and another will be associated with another famous companion *Ibn Abbas*, and the algorithm will attempt to correlate these two entries and might reveal some kind of associations and patterns that might highlight exegetical talents of both.

9.3.2.2 Discover patterns and correlations among the seven readings of the Qur'an

(Shah 2004) described the book authored by *Ibn Mujahid* on the seven readings of the Qur'an. Early scholars exercised extensive effort to preserve a large corpus of Qur'anic readings by early readers and grammarians. These readings specify subtle morphological and grammatical differences for the same Qur'anic

word. These differences has implications on the exegesis and legislative rulings, as is the case with verse 5:6

وَامْسَحُوا بِرُءُوسِكُمْ وَأَرْجُلَكُمْ إِلَى الْكَعْبَيْنِ
wamsahu bi rouwsekum wa arjulakum ilal ka'bain
and wipe / your heads / and your feet / till / the ankles
and wipe over your heads and wash your feet to the ankles

Qur'anic Verse -5:6

The Qur'anic reading scholars: Nafi', Ibn Amer, Hafs, al-Kesae and Yaqub recited '*arjulakum'* as accusative suggesting that foot should be washed, whereas scholars: Ibn Katheer, Hamza, Abi Amro and Aasem recited it as '*arjulikum'* as genitive suggesting that foot should be wiped.

Machine Learning algorithms can again be of assistance in discovering associations, patterns and interesting correlations between these subtle differences. The dataset can encode features as attributes such as case ending differences, substitution of a letter with other, singular and plural differences, morphological differences, etc.

9.3.2.3 Machine Learning on Syntactic and Linguistic Patterns:

Arabic and Linguistic scholars since early days investigated the Qur'an in search of linguistic phenomenon and interesting associations and patterns. Their investigation of such phenomenon in the Qur'an used to be done manually through laborious search in the text. As we saw in this paper, a researcher can define features of interest and prepare the data set accordingly, and let Machine Learning algorithms find associations and define interesting rules, which are then studied and verified by Qur'anic scholars further. In this way much human labour is saved and wider range of features are easily investigated.

For example, (Omar 2001), investigated the broken plurals of a singular noun in the Qur'an . He prepared a list of the occurrences of all such plurals -which counted 57 singular nouns- and manually tried to investigate the context of such plurals in attempt to discover certain patterns. For example, he found that the *jumu' al-qillah* (plural of fewness) usually comes in the context when fewness as

opposed to abundance is highlighted. In the presence of a linguistically tagged corpus of the Qur'an -like the Qur'anic Arabic Corpus- a dataset can be prepared with features as words surrounding the broken plural and their linguistic tags.

Computational linguistics can automate laborious linguistic search process for human researchers. When a tagged corpus is available at hand which tags parts-of-speech, grammar case, mode and other relevant morphological features for each word of the Qur'an, then it becomes easy to build customized queries for the computer to fetch. For example, grammatical shifts and anomalies in the Qur'an can be easily detected and their contexts can be observed prior to proper analysis. When it comes to empirical analysis of certain linguistic phenomenon, then linguistically-informed search can be of great help. (Abdul-Roaf, 2003) for example, tries to search for macro and micro logical coherence in Qur'anic discourse and notes that (p.75) "An exhaustive account of these textual features can be realised through the employment of modern text linguistic theory which is concerned with the analysis of texts and their lexico-grammatical and phonetic features ." To this end modern computational linguistics achieved considerable progress in automatic recognition of many such lexico-grammatical features.

9.3.2.4 Machine Translation

Machine Translation is an active field of research within Machine Learning and Computational Linguistics. Translating the Qur'an to world languages has also been an area of much scholarly research. For example, (Branca 2003) recognised the usefulness of computer assisted analysis when comparing several versions of existing translations and analyzing the differences .

Machine Translation has recently achieved good progress with the availability of big corpora of texts in both source and target language. Machine Learning can then employ statistical measures to suggest most likely translation of a word from source to target language based on their usage in these languages.

Translation of the Qur'an poses especial difficulty given the inimitability of the Qur'anic text where the source is impossible to be presented in a target language without losing some information in the process. Thus perfect translation of the Qur'an is almost impossible, and hence statistical machine translation can always

give predictions of best translations where the human researcher needs to make the final decision.

9.3.2.5 Computational Stylistics

Two factors that characterize a text are content and style. According to Muslims, the Qur'an is considered miraculous in both these aspects. Style of a text can be analysed through a set of measurable patterns called style markers. Research in computational stylistics includes software for genre detection and authorship attribution. Two main computational tasks involved are the extraction of the style markers and classification of the text according to these markers.

Popular style markers focus on distributional lexical measures at token level (e.g., word count, sentence count, character per word count, punctuation marks count, etc.), syntactic annotation (e.g., passive count, nominalization count, frequency of certain POS, etc.), vocabulary richness (e.g., type-token ratio, count of words occurring once, i.e., hapax legomena, count of words occurring twice, i.e., dis legomena, etc.) or counting frequency of some common words or function words.

When it comes to computationally analyzing the stylistics of the Qur'an, a specialised set of such style markers can be designed as the feature set for Machine Learning. For example, (Abdul-Roaf 2001) describes a number of syntactic features in the Qur'an which can be analysed computationally like: chandelier structures, multi-tiered structures, long argumentative structures, information listing structures, obligations, conditional clauses, tail-head/head-tail structures, hysteron and proteron, ellipsis, shifts, lexical repetition, recursive ties, phrasal ties and Qur'anic oaths.

9.3.2.6 Subjectivity Analysis

As the Qur'an is a book of guidance, its rhetorical instruments are employed in instructions, reminders, warnings, glad tidings, etc. (Jomier 1997) points out that 'the feature of argument and persuasion appears continually in the Qur'an. It is rare that long passages do not include, as parenthetical phrases, an interrogation, an apologetic allusion, an exhortation or a rebuke'.

Computational subjectivity analysis aims at computationally analysing the subjectivity of a text . Sometimes this research is referenced in literature as ‘sentiment analysis’ or ‘opinion mining’. Although, this research was motivated towards automatically analysing user opinions towards certain products over the internet, this research can be geared towards investigating subjectivity and rhetoric structure in the Qur’an.

One of the features we found distinguishing between Meccan and Medinan chapters is the *Qasas* or stories of past prophets repeated in Meccan chapters. As (Zebiri 2003) notes (p.111) ‘It is clear from the manner of their [stories] telling that the primary aim of these stories is not in fact historical. For example, little or no attention is paid to specifics such as date and place, and when lists of prophets are given they are not necessarily in historical order’ . Therefore, these stories can be studied computationally by extracting rhetorically interesting features and employing Machine Learning algorithms to discover interesting patterns and associations.

Bibliography

Abbas, N., Atwell, E. (2012). Qur'any: how to search for concepts rather than words in a corpus. Proc IVACS'2012, Leeds, UK.

Abdulbaqi, M. (1955) Al-mu'jam al-mufahras li al-fadh al-Qur'an (Aphabatical Index of the Qur'anic Words) (In Arabic). Dar al-adhami, Beirut. Available online at: http://www.Qurancomplex.com

Abdul-Roaf, Hussein (2003) "Conceptual and Textual Chaining in Qur'anic Discourse'. Journal of Qur'anic studies (5):2, 2003. pp:72-94.

Abdul-Raof, H. (2001). Qur'an Translation: Discourse, Texture and Exegesis. Routledge

Al-Rabiah, M. (2012) Reference Guide for the King Saud University Corpus of Classical Arabic. Available online at http://ksucorpus.ksu.edu.sa/?page_id=18

Al-Zubaidy, M. (died 1790 CE). Taj Al-A'rous [تاج العروس]. Published by Kuwait Government Press in 1973.

Ambros, A. A. (1987). Eine lexikostatistik des verbs im Koran. Wiener

Zeitschrift f¨ur die Kunde des Morgenlandes 77, 9–36.

Aston, G. and L. Burnard, (1998). The BNC Handbook: Exploring the British National Corpus with SARA, Edinburgh University Press.

Aone, C. and Bennett,S.W. (1994) Discourse tagging tool and discourse-tagged multilingual corpora. In Proceedings of the International Workshop on Sharable Natural Language Resources, pages 71–77, Nara, Japan.

Atwell, ES; Brierley, C; Dukes, K; Sawalha, M; Sharaf, A (2011) An Artificial Intelligence Approach to Arabic and Islamic Content on the Internet in: Proceedings of NITS 3rd National Information Technology Symposium. Riyadh 2011.

Atwell, E., Dukes, K., Muhammad, AB, Habash, N. et al.(2010) Understanding the Qur'an: A new Grand Challenge for Computer Science and Artificial Intelligence. Grand Challenges for Computing Research (2010). British Computer Society Workshop. Edinburgh

Baker, C., C. Fillmore and J. Lowe (1998). "The Berkeley Framenet project." In Proceedings of the 17th International conference on Computational Linguistics. ACL. NJ, USA.

Banerjee, S., and Pedersen, T. (2002) An adapted lesk algorithm for word sense disambiguation using WordNet. In CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, pp. 136–145, London: Springer Verlag.

Baldwin, B. (1997) CogNIAC: High precision coreference with limited knowledge and linguistic resources. In R. Mitkov and B. Boguraev, editors, Operational factors in practical, robust anaphora resolution for unrestricted texts,pages 38 – 45.

Barbu, C. (2003). Bilingual Pronoun Resolution: Experiments in English and French. PhD Thesis. University of Wolverhampton.

Bazargan, M. (1976). *Sayr-I tahawwul-I Qur'an*, (History of Quranic chronology). Vol I, Tehran, Qalam, 1976.

Beesley, K. R. and Karttunen, L. (2003). Finite-State Morphology: Xerox Tools and Techniques. Stanford: CSLI.

Bernard, J. (1986) The Macquarie Thesaurus. Sidney, Australia: Macquarie Library.

Black, W. Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., and Pease, A., (2006) Introducing the Arabic WordNet project. In Proc. of GWC-06, pages 295–299.

Branca, Paola (2003) "The translation of the Qur'an: A comparative approach based on a computer-aided analysis (Case study: Italian translations)" Journal of Qur'anic studies (5):1 2003, pp. 35-46

Brennan, S.E., Friedmann, M.W. and Pollard, C.J. (1987) A Centering approach to pronouns. In Proceedings of the 25th Annual Meeting of the ACL, pages 155–162, Stanford.

Brierley, C., Sawalha, M. and Atwell, E. (2012). Open-Source boundary-annotated corpus for Arabic speech and language processing. LREC 2012, Istanbul.

Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer version 1.0. Linguistic Data Consortium, University of Pennsylvania.

Budanitsky, A., and Hirst, G. (2006) Evaluating WordNet-based measures of semantic distance.Computational Linguistics 32(1): 13–47.

Chou´emi, M. (1966). Le verbe dans la Coran. Paris: Klincksieck.

Dolan, W.; Quirk, C.; and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In Proceedings of the 20th International Conference on Computational Linguistics.

Dror, J., Shaharabani, D., Talmon, R., and Wintner, S. (2004) Morphological Analysis of the Qur'an. Journal of Literary and Linguistic Computing 19(4): 431-452.

Dukes, K. and T. Buckwalter, T. (2010). A Dependency Treebank of the Qur'an using Traditional Arabic Grammar. In Proceedings of the 7th International Conference on Informatics and Systems (INFOS). Cairo, Egypt.

Dukes, K. and Habash, N. (2010). Morphological Annotation of Qur'anic Arabic. Language Resources and Evaluation Conference (LREC). Valletta, Malta.

Dukes, K., Atwell, E. and Sharaf, A.M. (2010). Syntactic Annotation Guidelines for the Qur'anic Arabic Dependency Treebank. Language Resources and Evaluation Conference (LREC). Valletta, Malta.

Fellbaum, C. (1998) WordNet an Electronic Lexical Database. Cambridge, MA: MIT Press.

Fillmore, C. (1976). "Frame Semantics and the nature of language." Annals of the New York Academy of Science.

Fillmore, C., C. Johnson and M. Petruck (2003). "Background to Framenet". Int. Journal of Lexicography, 16(3), 235-250.

Finkelstein, L., Gabrilovich, E.,Matias, Y., Rivlin, E., Solan, Z.,Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. ACM Transactions on Information Systems, 20(1), 116–131

Fiteih, M. (1983) Prepositions and Prepositional Verbs in Classical Arabic. PhD thesis, University of Leeds.

Gabrilovich, E., and Markovitch, S. (2007) Computing semantic relatedness using Wikipedia based explicit semantic analysis. In Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1606–11, Hyderabad, India.

Ge, N., Hale,J. and Charniak, E. (1998). A Statistical Approach to Anaphora Resolution. In Proceedings of the SixthWorkshop on Very Large Corpora, COLING-ACL '98, pages 161 – 170, Montreal, Canada.

Geissinger, Aisha (2004) "The Exegetical Traditions of Aisha: Some Notes on their Impact and Significance". Journal of Qur'anic Studies 6: 1 (2004), pp 1-20

Grosz, B., Joshi, A. K., and Weinstein, S. (1995). "Centering: A Framework for Modelling the Local Coherence of Discourse", NSF Science and Technology Center for Research in Cognitive Science. IRCS Report No. 95-01.

Gurevych, I., and Strube, M. (2004) Semantic similarity applied to spoken dialogue summarization In The 22nd International Conference on Computational Linguistics (COLING), pp. 764–70, Geneva, Switzerland.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten. I. H. 2009. The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11, 1 (November 2009), 10-18.

Hardie, A (forthcoming) "CQPweb - combining power, flexibility and usability in a corpus analysis tool".

Hirst, G., and St-Onge, D. (1998) Lexical chains as representation of context for the detection and correction malapropisms. In Christiane Fellbaum (ed.), WordNet: An Electronic LexicalDatabase and Some of Its Applications, pp. 305–332. Cambridge, MA: The MIT Press.

Hobbs, J. (1976) Pronoun resolution. Research report 76-1, City College, City University of New York.

Huang, S.; Graff, D.; and Doddington, G. (2002) Multiple-Translation Chinese Corpus. Linguistic Data Consortium, Philadelphia

Ibn-Katheer (d. 1372 C.E). Arabic Title: (تفسير القرآن العظيم) Tafseer Al-Qur'an. Dar Taibah, 1st Ed. 2008.

Jarmasz, M., and Szpakowicz, S. (2003) Roget's thesaurus and semantic similarity. In Proceedings of Recent Advances in Natural Language Processing, pp. 111–20.

Jiang, J. J., and Conrath, D. W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the 10th International Conference on Research in Computational Linguistics, Taipei, Taiwan.

Jomier. J (1997) The Great Themes of the Qur'an. London: SCM Press

Kennedy, C. and Boguraev, B. (1996.) Anaphora for everyone: pronominal anaphora resolution without a parser. In Proceedings of the 16th International Conference on Computational Linguistics (COLING'96), pages 113–118, Copenhagen, Denmark.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. Computational Linguistics. Vo. 29, Number 3.

Lappin. S. and Leass, H.J. (1994) An algorithm for Pronominal Anaphora Resolution. Computational Linguistics, 20(4):535 – 562.

Leacock, C., and Chodorow, M. (1998) Combining local context and WordNet similarity for word sense identification. In WordNet: An Electronic Lexical Database, pp. 265–83. Cambridge, MA: MIT Press.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the SIGDOC Conference 1986.

Li, Y.; McLean, D.; Bandar, Z.A.; O'Shea, J.D.; Crockett, K.; (2006) Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering. vol. 18, no. 8, pp. 1138-1150

Lin, D. (1998). An information-theoretic definition of similarity. In Proceedings of International Conference on Machine Learning, pp. 296–304. Madison, WI.

Maamouri, M., Bies, A. and Buckwalter, T. (2004). The Penn Arabic treebank: Building a large-scale annotated Arabic corpus. In NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt.

Manning, C. D., Raghavan, P. and Schütze, H. Introduction to Information Retrieval, Cambridge University Press. 2008.

McEnery, T and Wilson. A (1996). Corpus Linguistics. Edinburgh University Press, Edinburgh.

McEnery, T. and Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge Press.

Mihalcea, R., and Moldovan, D. I. (1999) A method for word sense disambiguation of unrestricted text. In Proceedings of the 37th Annual Meeting of

the Association for Computational Linguistics, pp. 152–8, Maryland, MD: Association for Computational Linguistics.

Miller, G., & Charles, W. (1991). Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1), 1–28.

Mitkov, R., Belguith, L. and Stys, M. (1998) Multilingual Anaphora Resolution. In The Third International Conference on Empirical Methods in Natural Language Processing, pages 7–16, Granada, Spain.

Moisl, H. (2009). Sura Length and lexical probability estimation in cluster analysis of the Qur'an. ACM Transactions on Asian language information processing (TALIP) 8, 4 (Dec. 2009), 1-19.

Muhammad, A. and Atwell, Eric, (2012a) "QurAna: corpus of the Qur'an annotated with pronominal anaphora", LREC 2012, Istanbul

Muhammad, A. and Atwell, Eric. (2012b) "QurSim: A corpus for evaluation of relatedness in short texts", LREC 2012, Istanbul

Muhammad, A. and Atwell, E. (2009) A Corpus-based computational model for knowledge representation of the Qur'an. 5th Corpus Linguistics Conference, Liverpool.

Oliver, I. (1993). Programming Classics: the world's best algorithms. Prentice Hall.

Omar, Ahmad Mokhtar (2001) "Multiple broken plurals of a singular noun in the holy Qur'an." Journal of Qur'anic studies (3): 2, 2001, pp. 170-199

Orasan, C. Evans, R. and Mitkov, R. (2000) Enhancing Preference-Based Anaphora Resolution with Genetic Algorithms. In Proceedings of Natural Language Processing - NLP2000, pages 185 – 195.Springer.

Patwardhan, S., and Pedersen, T. (2006) Using WordNet based context vectors to estimate the semantic relatedness of concepts. In Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together, pp. 1–8, Trento, Italy: Association for Computational Linguistics.

Patwardhan, S., Banerjee, S., and Pedersen, T. (2003) Using measures of semantic relatedness for word sense disambiguation. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, pp. 241–57, Mexico City, Mexico.

Pickthall, M. (1973). The meaning of the glorious Qur'an. Islamic Call Society.

Rabbani, M. S (2012). Al-Makki wal Madani, The Meccan and Medinan Chapters. Research Paper available at http://www.Qurancomplex.com/Tree.asp?section=2&TabID=2&SubItemID=1&l=arb&SecOrder=2&SubSecOrder=1

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989) Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics 19(1): 17–30.

Resnik, P. (1995) Using information content to evaluate semantic similarity. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448–53, Montreal, Canada.

Roget, P. (1962) Roget's International Thesaurus, 3rd ed. L. V. Berrey, and G. Carruth (eds.), New York: Thomas Y. Crowell Co.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. Communications of the ACM, 8(10), 627–633

Ruppenhofer, J., M. Ellsworth, M. Petruck, and C. Johnson (2005). “FrameNet: Theory and Practice.

Sadeghi, B. (2011). “The Chronology of the Qur’an: A stylometric Research Program.” Arabica. Vol. 58. 2011, pp: 210-299

Sawalha, M., Brierley, C., and Atwell, E. (2012). Predicting phrase breaks in classical and modern standard Arabic. LREC 2012, Istanbul.

Sayoud, H. (2012). “Author discrimination between the Holy Quran and Prophet’s statements”. Literary & Linguistic Computing. Vol 27(4): 427-444.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Comp. Survey, 34(1):1-47.

*SEM Shared Task (2013) Preliminary Announcement. Available at: http://www.aclweb.org/portal/content/sem-shared-task-2013-preliminary-announcement

Shah, Mustafa, (2004) "The Early Arabic Grammarians' Contributions to the Collection and Authentication of Qur'anic Readings: The Prelude to Ibn Mujahid's Kitab al-Saba" ". Journal of Qur’anic Studies (6):1, 2004

Shanqity, M. (d. 1973). Arabic title: (أضواء البيان) Dar Aalam al-Fawaed. 1st Ed. 2005

Smrz, O. and Hajic, J. (2006). The Other Arabic Treebank: Prague Dependencies and Functions. In Arabic Computational Linguistics: Current Implementations, CSLI Publications.

Taha, S. (2008). Tajweed Qur'an. Dar Al-Ma'rifah. Damascus

Thabet, N. (2005). Understanding the thematic structure of the Qur'an: an exploratory multivariate approach. Proceedings of the annual meeting of Association of Computational Linguistics. Pp. 7 – 12.

Washtell, J. (2011) "Compositional Expectation: A Purely Distributional Model of Compositional Semantics", in Proceedings of the 2011 International Conference on Computational Semantics (IWCS'11)

Weil, G. (1895). "An Introduction to the Quran. III" (translated by Frank Sanders, et al.), The Biblical World, 5/5. May 1895, pp. 343-59

Winograd, T. (1972). Understanding natural language. Cognitive psychology, 3:1–191.

Wright, W. (1967) A Grammar of the Arabic Language. Cambridge University Press. 3rd Ed.

Wu, Z., and Palmer, M. (1994) Verb semantics and lexical selection. In 32nd Annual Meeting of the ACL, pp. 133–8, Las Cruces, Mexico: Association for Computational Linguistics.

Yang, D., and Powers, D. M. W. (2006) Verb similarity on the taxonomy of WordNet. In Proceedings of the Third International WordNet Conference (GWC-06), pp. 121–8, Jeju Island, Korea.

Zebiri, Kate (2003) ‘Towards a Rhetorical Criticism of the Qur’an’, Journal of Qur’anic Studies 5:2 (2003), pp. 95-120.

Zesch, T., Gurevych, I., and M¨uhlh¨auser, M. (2007). Comparing Wikipedia and german wordnet by evaluating semantic relatedness on multiple datasets. In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pp. 205–8. Rochester, NY: Association for Computational Linguistics.

Zesch, T. and Gurevych, I. (2009). Wisdom of crowds versus wisdom of linguists – measuring the semantic relatedness of words. Journal of Natural Language Engineering 16(1): 25-59.

List of Abbreviations

BNC – British National Corpus

CA – Classical Arabic

FE – Frame Elements

LU – Lexical Units

MSA – Modern Standard Arabic

MUC – Machine Understanding Conference

NLP – Natural Language Processing

POS – Part-of-speech

QAC – Qur'anic Arabic Corpus

QADT – Qur'anic Arabic Dependency Treebank

QurAna – Corpus of Qur'anic Anaphora Tagging (Chapter 5)

QurSim – Corpus of Qur'anic Similar/Related Verses (Chapter 6)

TF-IDF – Term Frequency, inverse document frequency

Appendix A
Header of the download file for QurSim Dataset

```
<!--
* PLEASE DO NOT REMOVE OR CHANGE THIS COPYRIGHT BLOCK
*======================================================================
*  Annotation of Qur'anic Pronouns (version 0.1)
*  Copyright (C) 2011 Abdul-Baquee Muhammad
*  License: GNU Public License
*
*  This annotation contains marking pronons with <pron> tags and
*   indentifying their antecedents as well as the concepts.
*  This work used QAC (http://corpus.Quran.com) for segmentation IDs
*  and POS tagging.
*
*  TERMS OF USE:
*
*  - Permission is granted to copy and distribute verbatim copies
*    of this file, but CHANGING IT IS NOT ALLOWED.
*
*  - This annotation can be used in any website or application,
*    provided its source (TextMiningTheQuran.com) is clearly
*    indicated.
*
*  - This copyright notice shall be included in all verbatim copies
*    of the text, and shall be reproduced appropriately in all works
*    derived from or containing substantial portion of this file.
*
*  Check updates at (http://TextMinigtheQuran.com)
```

```
-->

<?xml version='1.0' encoding='utf-8' ?>

<chapter id='1'><verse id='1'><seg id='1'> بِ </seg>

<seg id='2'> سْمِ </seg>

<seg id='3'> ٱللهِ </seg>

<seg id='4'> ٱل </seg>

<seg id='5'> رَّحْمَٰنِ </seg>

<seg id='6'> ٱل </seg>

<seg id='7'> رَّحِيمِ </seg>

</verse><verse id='2'><seg id='8'> ٱلْ </seg>

<seg id='9'> حَمْدُ </seg>

<seg id='10'> لِ </seg>

<seg id='11'> لَّـهِ </seg>

<seg id='12'> رَبِّ </seg>

<seg id='13'> ٱلْ </seg>

<seg id='14'> عَٰلَمِينَ </seg>

</verse><verse id='3'><seg id='15'> ٱل </seg>

<seg id='16'> رَّحْمَٰنِ </seg>

<seg id='17'> ٱل </seg>

<seg id='18'> رَّحِيمِ </seg>
```

Appendix B
Header of the download file for QurSim Dataset

```
<!--
* PLEASE DO NOT REMOVE OR CHANGE THIS COPYRIGHT BLOCK
*========================================================================
*  Dataset on Qur'anic Verse Relatedness from Tafsir Ibn Katheer (version 0.1)
*  Copyright (C) 2011 Abdul-Baquee Muhammad
*  License: GNU Public License
*
*  This dataset lists pairs of verses that have been identified by Ibn
*  Kathir in his Tafsir book. After collecting these pairs, two further
*  passes were made manually to brand degree of relatedness.
*
*  Level '0':
*  seems very loosely related and should be understood by looking
*  into the context in the tafsir book.
*
*      Level '1':
*      These pairs are understandable by Human reader to be related, but
*      still might be difficult for training learning algorithms
*
*      Level '2':
*      These pairs are very much related and might be suitable for taining
*      machine learning algorithms.
*
*  TERMS OF USE:
*
*  - Permission is granted to copy and distribute verbatim copies
```

```
*    of this file, but CHANGING IT IS NOT ALLOWED.
*
*  - This annotation can be used in any website or application,
*    provided its source (TextMiningTheQuran.com) is clearly
*    indicated.
*
*  - This copyright notice shall be included in all verbatim copies
*    of the text, and shall be reproduced appropriately in all works
*    derived from or containing substantial portion of this file.
*
*  Check updates at (http://TextMiningtheQuran.com)
*
*  USAGE:
*
*      "uid"         : incremantal ID
*      "ss"          : source chapter number
*      "sv"          : source verse number
*      "ts"          : target chapter number
*      "tv"          : target verse number
*      "common"      : the number of common root words between the two verses
*      "relevance"   : the degree of relatedness as explained above
-->
<pma_xml_export version="1.0">
    <database name="related-verses">
        <!-- Table kathir -->
        <table name="kathir">
            <column name="uid">1</column>
            <column name="ss">1</column>
            <column name="sv">1</column>
```

```
    <column name="ts">1</column>

    <column name="tv">2</column>

    <column name="common">0</column>

    <column name="relevance">2</column>

</table>
```