FIRST YEAR TRANSFER REPORT

The Qur'an Annotation for Text Mining

School of Computing

By
Abdul-Baquee M. Sharaf
Supervisor
Dr. Eric Atwell

December 2009

CONTENTS

CHAPTER 1 – INTRODUCTION

This report presents my PhD research plan and provides the material for transfer review meeting. My research project aims at the discovery and extraction of interesting, non-trivial knowledge from the Qur'an using text mining techniques. There has been lots of research on the topic of text mining, mainly from biomedical field. The primary goal has been to retrieve knowledge that is hidden in text, and to present the filtered knowledge to users in a concise form. Text mining differs from data mining in the nature of the target data. While data mining investigated "structured" data residing in databases, text mining intends to mine for knowledge hidden in "unstructured" data within free texts. Text mining thus adds a layer of complexity in making the unstructured text ready for data mining algorithms. To this end, one of the challenging and laborious tasks has been to annotate the text under investigation with various linguistic information prior to applying machine learning and text mining algorithms.

To our knowledge, no academic research exists on text mining the Qur'an, and our objective is to pioneer a research enriching the raw Qur'an with linguistic and conceptual annotation and use it to "mine" for interesting knowledge hidden in the Qur'an.

In this introduction, I will first elaborate on my motivation and rationale for choosing the Qur'an as my candidate text for mining, and then I will present the scope of this project. Next I will highlight the novelty and original contribution of this research. Finally, I will outline the structure of the rest of this report.

1.1 – MOTIVATION AND BACKGROUND

The Qur'an is the religious text for Muslims. According to Muslims it is the words of God revealed on the prophet Muhammad 1443 years ago in classical Arabic language. The Qur'an claims to contain valuable information and gives answer and solution to many problems facing mankind. It also claims to be free from contradictions and discrepancies. In various places, the Qur'an challenged the mankind to author a book or a chapter of a book that would resemble the Qur'an in content and style. The Qur'an contains around 77,000 words which are organized into 114 varying size chapters called *surahs*. Each chapter in turn contains varying size verses or *ayat* (in total over 6200 verses). The Qur'anic scholars in the past fifteen centuries have been authoring books highlighting various linguistic, stylistic, scientific, rhetorical, and many hidden discoveries from the Qur'an in various other fields. Obviously, these scholars have been relying on their personal knowledge and familiarity of the Qur'an as no computational tools were available then. The Qur'an is characterized by holding vast information in unstructured and scattered –yet conceptually related- verses.

All these features make the Qur'an an attractive target for computational text mining with the objective of finding new information from the Qur'an in terms of hidden trends, relationships, patterns, coincidences and associations. Moreover, a historical analysis of the chronology of Qur'anic revelation gives more significance to the task of text mining from the Qur'an.

HISTORICAL BACKGROUND

The Qur'an was revealed over 23 years of the lifetime of the prophet Muhammad who was 40 years old when the first verse was revealed in the city of Makkah. The prophet started to preach these verses; initially secretly and later openly. The overarching theme of the Qur'anic verses is the emphasis on the monotheism doctrine in worshiping only one God and opposing the general polytheistic belief of the people of Makkah. Hence, majority opposed this new religion, but gradually adherents increased. The mission of Muhammad continued 13 years in Makkah. The general theme of the verses and chapters revealed in Makkah thus emphasized in establishing logical proofs for

monotheism detailing on the attributes of God and His supreme power. This is illustrated often by relating stories of past people and prophets and what happened to them when they rejected the monotheistic message. In general Makkah surahs emphasized the establishment of the monotheism doctrine, the prophethood of Muhammad and the reality of the Day of Judgment. As the people of Makkah were masters of the classical Arabic language, Prophet Muhammad considered this Qur'an – being the words of God- a miracle and challenged the Arabs of Makkah to bring a similar chapter like the Qur'an. Thus, the chapters of Makkah –in addition to the monotheistic tone- had a secondary objective of being literal, rhetorical and linguistic challenge for Arabs. This is evident is selection of strong words, phrases and Arabic constructs in Makki verses.

Later, the prophet migrated to Medina where the people welcomed him and allowed Islam to rule that city. Verses revealed in Medinah started to lay down Islamic law and jurisprudence, in addition to the continuing theme of Islamic monotheism. Medina period witnessed many battles of Islam and eventually Islam expanded to other nearby cities and tribes. In general, Medina surahs emphasized on establishing Islamic laws, ethics, morals, marital and family laws, monetary transactions, and relationship of Islam and Muslims with other world religions.

The above historical background gives us motivation to employ machine learning and text mining techniques to analyze the Qur'an in search for trends, patterns, association and distinguishing relationships in between and within those chapters revealed in Makkah and those revealed in Medinah. Moreover, these techniques can be used to extract feature sets for each chapters of the Qur'an which together can represent a unique signature for the Qur'anic authorship that might give some insight into the unique nature of the language of God. Chapter 5.3 discusses in detail some experiments I carried out exploiting some of these features to distinguish between Makki and Madani chapters.

RATIONALE

Motivated by the facts elaborated above, and after reviewing the literature and existing tools and techniques, we can rationalize further this project in the following points.

There has been increasing interests in Arabic NLP research. The first step towards such research is the preparation of a text annotated with various layers of linguistic features. The rich annotation and resources developed by this project will enable further research on the Qur'an or in classical Arabic in general.

The Qur'an being around 77,000 tokens (and around 15,000 types) is small in size by statistical analysis standards (Church and Mercer 1993). However, it is typical for many text mining and corpus linguistics projects to analyse works of a particular author. For example, (Plaisant et al 2006) mined for erotic expressions in Emily Dickenson's correspondence. (Stubbs 2005) used corpus linguistics in quantitative stylistics on Conrad. (Starcke 2006) analyzed Jane Austen and (Mahlberg 2007) focused on Dickens. See chapter 2 for more works.

The relative small size of the Qur'an rationalizes the feasibility of the annotation task. Moreover, repetitive constructs in the Qur'an and the nature of extensive rule based traditional Arabic grammar will enable semi-automated annotation and thus make it feasible to manage within a PhD scope.

There are considerable computational works on text analysis of the Bible. For example, University of Stellenbosch hosted the Association Internationale Bible et Informatique conference (AIBI-6) in the year 2000 under the theme of "from Alpha to Byte" (Cook 2002) bringing together many computational research on Bible. Refer to chapter 2 for more details. These bible researches rationalize initiating similar research on the Qur'an which shares the same genre.

The Qur'an being a book extensively studied over the past 15 centuries is characterized by the availability of a huge library of scholarly works on Qur'anic exegesis, literary analysis and linguistic commentaries. These works will provide useful material in preparing annotation guidelines, learning algorithms and in evaluating text mining results.

Traditional Arabic grammar is characterized by a large set of rules that caters well for various complexities of the language. Moreover, the Qur'an has been exhaustively analysed syntactically – albeit not in computational format. These rules and analysis rationalize for learning classifiers for the Qur'an and classical Arabic.

The field of machine learning and text mining have seen many off-the-shelf classifiers which can save much effort in developing algorithms and allow more effort to prepare and enrich the raw Qur'an text as input to these algorithms.

1.2 – Problem Statement

This research project aims at creating an enabling computational environment for text mining the Qur'an. This involves defining the text mining objectives, and based on that annotating the raw Qur'an with linguistic and domain information. Using this annotated corpus, various text mining tasks including: information extraction, text categorization, concept linkage and discovery of associations and patterns are tested and evaluated.

The project adopts CRoss Industry Standard Process for Data Mining or CRISP-DM (Shearer 2000) methodology, and considers investigating chapters 2 and 7 of the Qur'an as a pilot project before focusing later on the entire Qur'an.

1.3 – Contribution and Novelty

The originality, contribution and novelty of this research can be considered in the following areas:

- Developing specialized corpus and resources for text mining the Qur'an in particular and NLP research on classical Arabic in general.
- Pioneering the research in the field of text mining the Qur'an by enriching the raw Qur'an with annotation and developing other necessary resources.
- Using various existing machine learning algorithms in a novel way for the first time in the domain of religious scripture written in classical Arabic – i.e., the Qur'an.

1.4 – Organization of this report

The rest of this report is organized as follows. Chapter 2 is a literature review investigating existing resources, tool and techniques both for the Qur'an and text mining. Chapter 3 gives detailed overview of the research plan following the six phases of the CRISP-DM methodology. Chapter 4 gives an overview of the work completed so far. Chapter 5, elaborates further my work progress through some computational exercises I carried out on the Qur'an. Finally, chapter 6 concludes the report highlighting on my contribution and rationalizing why I should proceed further in this research and get a 'green light' in this review meeting.

CHAPTER 2- LITERATURE REVIEW

2.1 QUR'ANIC RESOURCES

The original Arabic Quran has been circulating in digital form online for some time. However, early copies used to suffer from errors mainly due to missing diacritics and lack of encoding techniques to incorporate special characters that match the printed Quran. Motivated by these shortcomings, there have been few projects to produce error-free machine readable Quran. Most famous among them is the Tanzil project [http://tanzil.info/] which allows downloading the Quranic text in various formats including XML and SQL dump. I used these resources in carrying out my experiments so far, see chapter 5 for details.

Searching the Quran has been the focus of many websites. However, in most cases searches were limited to Keyword search as in http://www.islamicity.com/QuranSearch/. A number of sites allow root word search as well, for example: http://tanzil.info/. Similarly, there have been few sites allowing proximity or Boolean search over English translation of the Quran, for example http://quod.lib.umich.edu/k/koran/.

There have been many early works by Quranic scholars of Tafsir (the scholarly comments on the Quran). A common trend followed by many scholars of Tafsir is to attempt understanding first a verse by relating it with other verses, then relating it to Haidth (saying of the Prophet Muhammad), then commentaries made by students of Prophet Muhammad (and their students), and finally analysing classical Arabic lexicons. Most notable books of Tafsir following this methodology are (At-Tabari 2000), (Al-Baghawi 1997) and (IbnKatheer 1999), which are also available online in machine readable form.

There are few books that only explain the difficult vocabulary of the Quran, for example (Ad-Dinawry 1978). Also, another category of books deal with the science of Quran and Tafsir where the focus is not exhaustively on every verse but to extract general rules that help understanding the Quran. (As-Soyouti 1987) and (As-Sabt 1996) are two such works.

Traditional Arabic grammar defines many syntactic rules that relate words and segments in a sentence. There has been many works in Arabic that exhaustively produces grammatical parsing for the Quran for example (Salih 2007),(Al-Kharrat 2007) and (Darwish 1999).

(Al-Qahtani 2005) gives an extensive categorization of modern standard Arabic verb valence based on Case Grammar (CG) as described by (Fillmore 1968). Based on the assumption that CG is adequate to classify all verbs of a language and is universal across languages, Al-Qahtani went on to specify valence according to Cook's Matrix Model (Cook 1979) and its extension that includes 24 cells. According to this matrix five cases (Agent, Experiencer, Benefactive, Object, Locative) are plotted horizontally and type of verb (State, Process, Action) vertically. The date was taken from 8327 verbs from a lexicon (Al-Qahtani 2003) and most frequent 200 verbs were exhaustively sorted to a cell in the matrix, and thus proved the suitability of Cook's model for Arabic valence.

(Fiteih 1983) studied the prepositional verbs considering the Quran as his corpus. He could classify four classes of Quranic 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 [2.1]), or a nominal and a prepositional object (e.g., send against someone something as in [2.2]), or two prepositional objects (e.g., come forth unto someone from some place as in [2.3]), or one

nominal object and two prepositional objects [2.4a] or one prepositional object and two nominal objects [2.4b].

[2.1] *And when he saw their hands **reached** not to it, he mistrusted them.. [11:70]*
[2.2] *For We **sent** against them a furious wind, [54:19 Yusuf Ali Translation]*
[2.3] *Then he **came forth** unto his people from the sanctuary [19:11]*
[2.4] *a. And Allah hath favoured some of you above others in provision [16:71]*
*b. He hath **bestowed** on those who strive a great reward above the sedentary[4:95]*

(Shamsan 1986) studies the transitivity and intransitivity of Quranic verbs. He analyzed the valences of these verbs and tried to link between the form of these verbs and the semantic significance. He also observed the shift of a verb from intransitive to transitive sense based on semantic characteristics.

(Mir 1989) observed that quite a lot verbs in the Quran are used in idiomatic sense rather than literal meaning of the verb. He went on to list such expressions in the Quran. Some examples are given in the following quote.

> When a man's "eyes become cool", it means that he is pleased. A person who "brings down his wing" for you is being kind to you, but if he "bites his fingers" at you, he holds you a severe grudge. If you think you lack the gift of fluent speech, you can pray to God to "untie the knot in your tongue" (Mir 1989: 2-3)

There is not much computational research on the Qur'an. Here I present only two academic researches I could find. (Thabet 2005) applied hierarchical cluster analysis to lexical frequency data abstracted from the Qur'an to see if this would yield a classification of the chapters which could be useful in understanding its thematic structure (particularly the Makki from Madani). The results were interesting, but were compromised by a problem caused by the large variation in the length of the chapters, which range from fewer than ten words to several thousand.

(Shenassa and Khalvandi 2008) evaluated the verbs in chapters 2 and 3 of five English translation of the Qur'an in comparison with original Arabic verb and checked against the number of similarity between translations and source in terms of verb process according to Halliday Grammar. They found that Yousuf Ali's translation is better than others.

BIBLE STUDIES

Computational tasks on the Bible can prove useful for the Qur'an as both might fall under the same genre. There are few projects on the Bible. (Resnik et al 1999) described an ongoing project to create parallel multilingual corpus for the 66-books of Bible tagged with information about *book*, *chapter* and *verse*. (Chew et al 2006) evaluated the Bible to be a resource for cross-lingual information retrieval. To measure similarity their retrieval system constructed a matrix of 31,104 by 31,104 for each language pair where each cell measured similarity between two languages. They retrained the system on the Arabic Qur'an and three other languages: English, Russian, and Spanish.

2.2 SUBJECTIVITY ANALYSIS

Recently there has been an eruption of research activities in the area of opinion and sentiment mining (Pang and Lee 2008) because of a surge of opinionated information available online. Opinion and sentiment mining falls under general "subjectivity analysis" of a text (Banfield 1982) as opposed to traditional objective fact-based analysis. Other elements that can fall under subjectivity analysis are

emotion and speculation. The Qur'an is a text that contains both objective fact-based statements (2.5 and 2.7 below from verse 98:5-6) as well as opinions either positive (segment 2.6 below) or negative (segment 2.8 below).

(2.5) <objective>And they were not commanded except to worship Allah , [being] sincere to Him in religion, inclining to truth, and to establish prayer and to give zakah.</objective>

(2.6) <subj-positive> And that is the correct religion.</subj-positive>

(2.7) <objective> Indeed, they who disbelieved among the People of the Scripture and the polytheists will be in the fire of Hell, abiding eternally therein. </objective>

(2.8) <subj-negative>Those are the worst of creatures </sub-negative> [98:5,6]

Opinion mining can be broken into the following subtasks (Popescu and Etzioni 2005; Esuli and Sebastiani 2006):

1. Identify product features
2. From a review text identify the sentences/phrases which are fact-based (objective) and those which has positive-negative polarity (subjective)
3. If a sentence/phrase is subjective, determine its polarity (positive or negative)
4. Determine the strength of the polarity (e.g., very positive, mild negative, etc.)

(Popescu and Etzioni 2005) had decomposed the problem of review mining into the four subtasks above: identify product features, identify opinions regarding product features, determine the polarity of opinions and ranking opinions based on their strength. Their system takes as input a product and a set of reviews, and outputs a set of product features, accompanied by a list of ranked associated opinions. They started with parsing the review texts by MINIPAR (Lin 1998) and applied simple pronoun resolution. Explicit features are extracted through an information extraction system called KnowItAll (Etzioni et al. 2005) which finds candidate facts through generic extraction patterns and assesses the results using PMI (Point-wise mutual information) on web searches. They also employ WordNet in distinguishing parts from properties while finding explicit features. Again, they use WordNet synonymy and antonymy information to group the adjectives in a set of initial clusters while finding implicit features. In the subtask of finding opinion phrases, they employ relaxation labelling method to find the semantic orientation of words in the context of given product features and sentences. Thus this unsupervised classifier takes a set of semantic orientation labels like {positive, negative, neutral} and a set of objects (like words) along with initial probabilities on associations and neighborhood influence, and the classifier then assigns globally labels to objects taking into account local constraints. Their system relies on various extraction rules.

I think this paper is rich in using a wide range of NLP and machine learning techniques and many Qur'an mining tasks can be approached in similar ways. This paper gives rationale for annotating the Qur'an with syntactic, named entity and pronoun resolution as well as creating a wordnet type of lexical resource for Qur'anic words.

(Esuli and Sebastiani 2006) described SentiWordNet, a lexical resource which quantified each wordnet synset into objective, positive and negative scores that sum up to 1.0. For example, the adjective [estimable (3)] in wordnet means "maybe computed or estimated" is always objective terms and has score of (objective=1.0, positive=0.0, negative=0.0), whereas [estimable(1)] meaning "deserving of respect or high regard" is given a score of (objective=0.25, positive =0.75, negative=0.0). This work

can inspire creating Wordnet type of hierarchy for the Qur'an as well as tagging each sense in with similar objective, positive and negative scores.

(Wiebe et al 2005) describes MPQA or Multi-Perspective Question Answering opinion corpus which contains manual annotation of over 15,000 subjective expressions in nearly 9,000 sentences with contextual polarity of positive, negative, both or neutral as expressed in examples below (subjective expression underlined and polarity in brackets).

(2.9) Thousands of coup supporters celebrated (positive) overnight, waving flags, blowing whistles . . .
(2.10) The criteria set by Rice are the following: the three countries in question are repressive (negative) and grave human rights violators (negative)
(2.11) Besides, politicians refer to good and evil (both) only for purposes of intimidation and exaggeration.
(2.12) Jerome says the hospital feels (neutral) no different than a hospital in the states.

Inter-annotator agreement for this corpus was 82% and kappa value of 0.72. The corpus is publicly available with annotation guideline and can help us in annotating subjective expressions in the Qur'an.

(Turney 2002) takes a review text, tags it through Brill POS tagger for adjectives and adverbs and then identifies the semantic orientation of the phrases in this text as positive or negative. Finally the given review is assigned a class as recommended (thumbs up) or not recommended (thumbs down) based on the average of the semantic orientations of the phrases. Semantic orientation of a phrase is calculated by PMI of that phrase with a reference word like "excellent" or "poor". After POS tagging, he extracted pharses of two words based on defined patterns involving mainly the pattern (JJ NN) like "low fees" or "online experience".

Various non-factual applications of text mining are shown in the following table.

category	application	Example paper
Opinion mining	Binary classification of a review as "thumbs up" or "thumbs down"	(Turney 2002)
Opinion mining	Determine whether a political speech is in support or opposing to the issue under debate	(Thomas et al 2006)
Opinion mining	Classify election forums by "likely to win" or "unlikely to win"	(Kim and Hovy 2007)
Sentiment analysis, agreement detection	Given a pair of text decide if they should receive same or different sentiment labels	(Snyder and Barzilay 2007) (Thomas et al 2006)
Subjectivity analysis	Decide if a given document contains subjective information or not.	(Godbole et al 2007) (Wiebe et al 2004)
Viewpoints and perspectives	Classifying a text as in support of some political view like liberals, conservative, etc.	(Mullen and Malouf 2006)
Viewpoints and perspectives	Placing texts along an ideological scale	(Laver et al 2003) (Martin and Vanberg 2008)
Emotion analysis	Classify a text into the six universal emotion categories (Ekman 1982): anger, disgust, fear, happiness, sadness, and surprise.	(Liu 2003) (Alm et al 2005) (Subasic and Huettner 2001)

Table 2.1 – examples of subjectivity analysis applications

2.3 Stylistics Analysis

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 analyzed through a set of measurable patterns called *style markers*. Works in computational stylistics exists 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., dislegomena, etc.) or counting frequency of some common words or function words.

(Holmes 1992) uses multivariate statistical approach to tackles the problem of author attribution. He measures the richness of vocabulary through five variables which are based on distribution of the frequency of tokens and types with relationship with text length. Among the variables is also analysis of hapax legomena and hapax dislegomena. He then applied this analysis on the Book of Mormon which Joseph Smith claims to be a translation of ancient God revelation and "Doctrines and Covenants" which also claims to be a direct revelation from God. These writings are compared with several of Joseph Smiths personal writings through against the defined five vocabulary richness variables. If the claims of Joseph Smith was correct, then these works should form separate clusters, but the analysis showed the opposite suggesting that Smith may have claimed his own writings to be a revelation from God. This stylometric analysis can be valuable for our project in distinguishing between the sayings of Prophet Muhammad and the Qur'an text, also in checking the stylistic difference between Makki and Madani chapters.

(Stamatatos et al 2000) describes a system for author attribution and genre detection tested on Greek corpus but can work in any application. They chose style markers that are adapted to the output of existing NLP tools like POS tagger (word, sentence, punctuation), and that which are adaptable to analysis-level of the NLP tool (like words remained unanalyzed after certain pass). They used multivariate text categorization tools.

2.4 Text Categorization

Automated text classification is the task of assigning a category to a document. This task can be "hard" where the system should make binary decision on choosing the right category, or it could be "soft" –or semi-automated- where the system only presents a list of ranked possibilities and a human makes the decision (Sebastiani 2002).

There are two main approaches towards text categorization: knowledge engineering approach and machine learning approach (Feldman and Sanger 2007).

Knowledge engineering approach towards text categorization is achieved through manual development of classification rules by a domain expert. An example of such a system is CONSTRUE system (Hayes 1992) for Reuters. The drawback of such approach is the labor intensive effort to develop such a knowledgebase.

In the machine learning approach, the classifier is built automatically by learning the properties of categories from a set of preclassified training set. (Sebastiani 2002) is a good survey of various machine learning classifiers and table 2.2 below gives a summary of best performance results among each category of classifiers.

Classification type	Results reported by	Score
Support vector machines	(Dumais et al. 1998)	0.920
Decision trees	(Dumais et al. 1998)	0.884
Committee of classifiers	(Weiss et al. 1999)	0.878
Example based	(Lam and Ho 1998)	0.860
Regression	(Yang 1999)	0.855
Bayesian net	(Dumais et al. 1998)	0.850
Neural networks	(Yang and Liu 1999)	0.838
On-line linear	(Degan et al. 1997)	0.833
Decision rules	(Cohen and Singer 1999)	0.827
Naïve Bayes probabilistic classifier	(Dumais et al. 1998)	0.815
Rocchio batch linear	(Joachims 1998)	0.799

Table 2.2 – Popular classifiers for text categorization and their performance

It is evident that SVM, boosting and regression methods perform the best. (Yang 1999) has found that various methods perform comparably when training set contains over 300 instances per category, but when the number of positive instance in training data are small (like 10 instances) then regression like methods like linear least-square or k-nearest neighbours are better.

(Blake and Pratt 2001) highlight the importance of selecting features that are semantically rich and thus informative and useful to the domain. They showed empirical results from medical domain. Selecting a feature considering the phrase "breast cancer" is richer semantically than choosing two separate features as "breast" and "cancer", because the former is a domain term which can be equated with "neoplasm of the breast". This work highlights the importance of domain knowledge in selecting semantically rich feature set prior to learning classifiers on these features. In our project we also need to find semantically rich set of features to represent Qur'an which branches from domain and ontological knowledge and not just linguistic features. These features will follow directly from the nature of text mining queries intended from Qur'an: a task we will carry during the pilot project.

2.5 Clustering

Clustering is an unsupervised task to group unlabelled text into meaningful clusters without prior information. Documents are usually represented as a vector of features –often bag of words- and the task is to cluster these documents based on some similarity function. (Dhillon et al 2002) shows that this task is very computationally expensive and only 5,000 webpage can contain a vocabulary size of 15,000 words, which is a bottleneck for most effective clustering algorithms like support vector machines or k-nearest neighbour, etc. (Hotho et al. 2003) shows that incorporating background knowledge –in their case WordNet synset- makes clustering more efficient with an increase of 10% in purity. (Bloehdorn and Hotho 2006) extended the work using boosting algorithms and incorporating medical (MeSH Tree structure Ontology) and agricultural ontology (AGROVOC). These work motivate us to create domain ontology for the Qur'an and my pilot project should give a good assessment on the feasibility of such task.

2.6 Pattern Discovery: Distributions, Frequent sets and Associations

One of the most "lucrative" promises of text mining is the discovery of hidden information. For example, in our context, we might want to see the proportions of verses in the Qur'an that describes the concept of "fighting" and their distribution across Makki and Madani chapters. We might want to narrow the focus and want to see a distribution of verses mentioning both "praying" and "fasting" together. When we restrict the frequency to a minimum support level then we are interested in frequent and near frequent sets (Feldman and Sanger 2007). Finding such frequent sets is a step towards fining associations. An association rule presents a relation between two concepts, for example, a link between the concepts of "fighting" in the Qur'an "Madani surah" is evident in the Qur'an, as the Prophet was not in position to fight while in Makkah. These probable association is investigated in terms of confidence and support, for example, we might say "80% of verses that mention 'prayer' also mention 'Zakat (giving alms) and 1.5% of all verses of the Qur'an mentions both prayer and Zakat. Machine learning can be employed for finding all association rules with a user-defined confidence and support. (Agrawal et al. 1993) is a seminal work in this field and focus on investigating market-basket type of association for example finding that "90% of transactions that purchase bread and butter also purchase milk". (Amir et al. 2003) introduces the concept of maximal association rules. It is motivated by phenomenon in text where "Linux" is associated with low confidence (wrongly) with "open source" because of high confidence of associating with "windows". They described algorithm to overcome this problem.

(Silberschatz and Tuzhilin 1996) gives a good presentation on the topic of "interestingness" of the patterns produced by data mining tools. It is noted that many patterns returned are not interesting to the human user. They emphasize on subjective view of interestingness that vary from one user to other in oppose to objective measures like the confidence and support. They considered a pattern interesting to a user for two reasons: a) when it is unexpected and hence 'surprising' and b) when the user can 'do' something with it. In our context this provides good material when we define the objective of our system and what the user community – mainly Qur'anic scholars – consider interesting from a text mining on the Qur'an.

2.7 Text Mining in Biology and Biomedicine

There is an overwhelming amount of biomedical knowledge in textual form that triggered lots of research in text mining this domain particularly. These researchers are eager to generate new hypothesis after discovering associations, patterns and trends from past knowledge (Swanson 1990). Text mining is being applied in areas such as finding functional relationship among genes, establishing functional annotations, discovering protein-protein interactions, interpreting array experiments, associating genes and phenotypes, etc (Ananiadou and McNaught 2006). This field has flourished through increasing number of dedicated workshops, tutorial, etc. A centre has been established in UK called NACTEM to provide text mining service to the academic community. Researches in this field are helpful for my project as both deal with a domain dependent text mining. Many lexical, terminological and ontological resources exist in this field (Bodenreider 2006) that might motivate building similar knowledge repositories for the Qur'an. There are several corpora available which are annotated with both linguistic as well as conceptual information with detailed annotation guidelines (Kim and Tsujii 2006) which will prove helpful in our project. Moreover being domain dependent, several biomedicine text mining system relied on rule-based information extraction (McNaught and Black 2006) which can perform well for the Qur'an mining as well. (Park and Kim 2006) presents researches in named entity recognition in biology which relied on various methods like dictionary-based, rule-based, machine learning, and hybrid methods.

CHAPTER 3 - DETAILED DESCRIPTION OF THE PROJECT

CRoss Industry Standard Process for Data Mining or CRISP-DM (Shearer 2000) is an application-neutral process model for data mining developed by industry leaders. Our plan is to adhere to these generic process guidelines in order to be on track by following best practices. This model organizes data mining projects into six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. Figure 3.1 below illustrates these steps and shows its iterative nature.

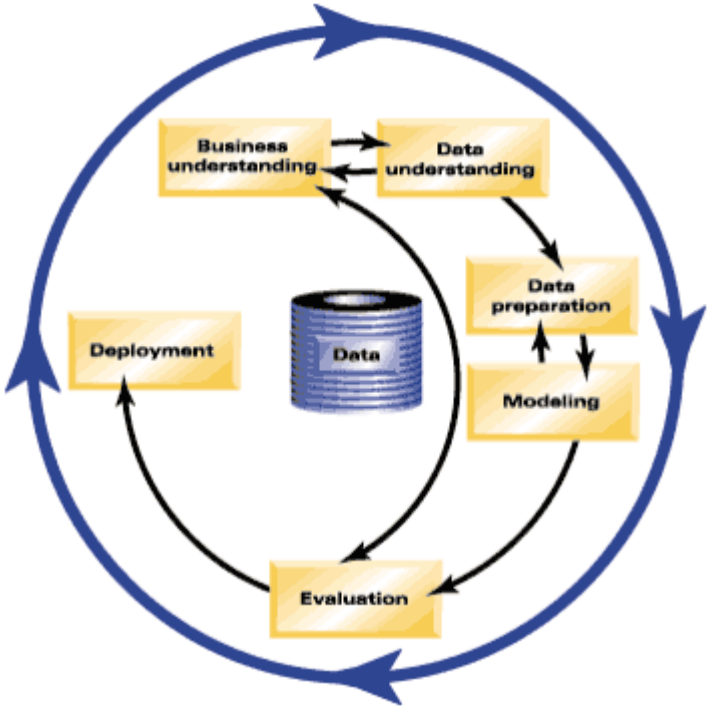


Figure 3.1 - CRISP-DM process Model (Shearer 2000)

Each of the six phases is further broken down into key steps, as shown in the figure 3.2 below.

Business Objectives	Data Understanding	Data Preparation	Analysis & Modelling	Evaluation	Reporting/ Deployment
Determine Business Objectives	Collect Initial Data	Select Data	Select Analysis/ modelling technique	Evaluate Results	Plan Deployment
Assess Situation	Describe Data	Clean Data	Generate Test Design	Review Process	Plan Monitoring & Maintainence
Determine Data-Mining Goals	Explore Data	Construct Data	Build Model	Determine next steps	Produce Final Report
Produce Project Plan	Verify Data Quality	Integrate Data	Assess Model		Review Project
		Format Data			

Figure 3.2 - Key steps under each phase of CRISP-DM

In what follows, we adopt this model for the purpose of our text mining project and describe each phases and steps accordingly.

3.1 Phase 1 – Business Understanding

This phase assesses the objectives of this project and the value it will add to our clients (i.e., mainly the Qur'anic scholars). Then translates these objectives into specific technical objectives in text mining terms, assess the available tools and resources and produce a detailed plan. Following are more elaboration on each of these steps.

3.1.1 Determine the Business Objective

In this step, we are interested in identifying our key clients who would benefit from this project. Next, identify what each of these clients want from this project – i.e., their key objectives and queries from our text mining system. What are the key success factors and how to measure each of these factors?

Who are the user community?

Our system should be beneficial to a wide range of user community. Following are the main category of users:

Qur'anic scholars: they are the main beneficiary of our system. They are familiar with the content of the Qur'an, and hence are generally not interested in knowing if certain piece of information is or is not in the Qur'an. Rather, they want to validate empirically many findings and claims made by early Qur'anic scholars on certain rules concerning the Qur'an. In general, they should feel very excited by a tool that enables them to carry empirical analysis on the Qur'an which would otherwise be very tedious or impossible by manual investigation.

Arabic Scholars: they are interested in Qur'an as being a text of classical Arabic. They want to search for Qur'anic evidence in support for some Arabic grammar rules. They want to read concordance lines for the usage of certain lexical terms and observe the behaviour of certain part-of-speech. In general their interest matches with the field of computational stylistics.

Qur'anic students: they are perhaps non-Arabs but have interest in the Qur'an both for content and language. They want to search for contents: either beyond keyword or without having the knowledge of proper keywords, and hence want to search by topics and concepts. They want to learn how various Qur'anic words are related linguistically to each other by relations like synonym, hyponym, metonym, etc.

What they want?

Each user category identified above would benefit from our system differently. Their expectations might be very high. However, we need to scale down the full set of client objectives into feasible and realistic objectives given the scope of a PhD research. Following table lists this shortlist of objectives per user category.

User	What they want?
Qur'anic Scholar	- Validate existing claims by early scholars through empirical methods - Link scattered but related verses - Find interesting correlations and association between verses
Arabic Scholar	- Qur'anic evidence of grammar rules - Linguistic behaviour of certain words or POS
Qur'anic Students	- Search for a topic or concept - Hierarchical relation of Qur'anic words

Table 3.1 – Key users and their objective from Qur'anic text mining solution

It is worth noting that this initial set of objectives needs to be revisited and amended after conducting the pilot study. Following are elaborations with examples on some of the objectives mentioned in the above table.

As the Qur'an has been a subject of extensive scholarly research by early Muslims, we have access to many observations, findings and claims based on manual investigation of the Qur'an. For example, finding distinctive characteristics that differentiates Makki verses from Madani verses has been a subject of interests by many early scholars. Some of them (Maimoun Ibn Mihran, see (As-Soyouti 1987 p. 52)) made a generic observation that whenever the construct "O mankind" is used in a verse, then this verse must be Makki, and whenever the construct "O you who believe" is used in a verse, then this verse must be a Madani . Another scholar made a claim that when certain chapter starts with some initial letters, then these letters are the most frequently used letters in that chapter, or words containing these letters in that chapter carries the theme of that chapter. Our text mining system will enable scholars for the first time to empirically validate these findings.

The Qur'an describes itself in the verse 39:23 as "a consistent Book wherein is reiteration". There are many verses –and segments within a verse - in the Qur'an that attach to the same concept or topic. For example, in 80 different scattered verses a command to perform prayer (i.e., the second pillar of Islam) has been followed by a command to pay Zakat (i.e., an annual expenditure to be paid for the needy, and the third pillar of Islam). Sometimes it is difficult even for Qur'anic experts to relate all these instances in memory. And an analytical tool would be of great help in this regard. Figures 3.7 and 3.11 in the following pages provide more appreciation of this feature.

The word "Ummah" was mentioned 49 times in the Qur'an. This word is commonly used in modern standard Arabic to mean a nation, like the term "al-Ummah al-Islamiyyah" meaning "the Islamic nation" or the term "al-umam al-muttahidah" meaning "the United Nations". However reading the concordance lines in the Qur'an, a researcher would notice that this word has three different meanings mostly to mean 'a nation', but also to mean 'a short period' as in verse 11:8, or a 'leader' as in verse 16:120. A wordnet type of hierarchy of Qur'anic words will be of great benefit both for learning purpose and other computational tasks.

WHAT ARE THE SUCCESS CRITERIA?

These objectives must be further justified by setting measurable success criteria. Any computational exercise should not be attempted if manual work can achieve the same result with minimum effort. Table 3.2 below is a list of success criteria and their measure for our project.

Success criteria	How to measure?
Discovered new correlation	Verified as OK by Qur'anic scholars
Related verses are linked	Precision and recall
Finding errors in early claims	Verified as OK by Qur'anic scholars
Positive feedback from users	Starred ratings
Accepted publications	How reputed is the conference or journal

Table 3.2 – how to measure success of this project

Following is more elaborations on some of the points mentioned in the table above.

Qur'an is a very complex text for machine understanding and various external and contextual factors need to be considered. Any results produced computationally showing apparent trends or correlations need to be verified by Qur'anic scholars.

Many books of Tafsir (Qur'anic commentary) when commenting on a verse, cite some of the other verses that are related. Taking this set of verses as a benchmark, our system can be evaluated by adding more related verses (i.e., more recall), which are verified to be correctly related by scholars (i.e., precision). Figure 3.11 depicts this concept where many scattered verses accumulatively contributes to cover a certain topic or concept.

With text mining tools, it is easy to verify certain claims like the one made earlier on criteria to differentiate Makki and Madani verses by the construct "O mankind". Chapter 2 which is a Madani chapter mentions twice (verse 21 and 168) the construct "O mankind", thus refuting the claim that this construct only occurs in Makki chapters.

An obvious measure of success of this project is the publication of paper in both Qur'anic studies and text mining conferences. It is our intention to present papers to following list of conferences and journals.

Potential publication	Possible topic
Corpus Linguistics 2011	Framework of text mining based on multi-layer annotation
LREC 2012	- Annotation guidelines - Tools and resources for text mining
EMNLP	Computational methods to link similar concepts in the Qur'an
ACL	Employment of text mining algorithms for Qur'an
Journal: Computer and Humanities	A computational framework for text mining a religious text
Journal: Language and Literature	Computational analysis of stylistics of the Qur'an
Qur'an and Technology Conference [Arabic]	A tool for text mining the Qur'an

Table 3.3 – Possible publication opportunities

3.1.2 Assess the Situation

In this step, we outline the resources, the assumptions, the risks and measures to mitigate them.

Resources

In terms of data, we have machine-readable Qur'an available in the public domain [*http://tanzil.info*] both as raw text and as XML format. We also have Haifa morphological and POS tagged corpus (Dror et al. 2004) which is available publicly with 80% accuracy. The Quranic Arabic Corpus (Dukes and Habash 2010) is also publicly available but is still undergoing manual verification. We also have a rich library of Qur'anic science and Arabic grammar books which will help in preparing the annotation guidelines process of the Qur'an.

In terms of tools, we have MMAX2 (Müller and Stube 2006) as a highly customizable tool for creating, browsing, visualizing and querying linguistic annotations on multiple levels. Weka (Witten and Frank 2005) is a popular machine learning toolkit for learning classifiers, clustering and visualizing results. NLTK provides good modular solution towards many text processing tasks. Chapter 5 details on some of my experiments with these tools.

In terms of methodology we follow the general phases, steps and templates provided by CRISP-DM guidelines. As for the machine learning classifiers for text mining, there are various type of algorithms like probabilistic, decision trees, decision rules, regression, linear, example-based, Bayesian net,

neural net, SVM or a committee of methods. Results show better performance of SVM classifiers, however, the decision of the particular algorithm for this project is deferred after the pilot project.

As for the human resource, apart from me being the main researcher and Dr. Atwell being my supervisor, this project may resort to the expertise of Dr. Hussein Abdul-Raof and Dr. Mustapha al-Sheikh from Arabic and Middle Eastern Studies at Leeds University for consultation. Based on my background, I feel comfortable working with this project, as I am fluent in Arabic and am very conversant with the Qur'an and books of Tafsir and have developed skills in Qur'anic science over the past 10 years. Moreover, I have six years industry experience in software quality assurance and ERP solutions with SAP, which I believe will contribute positively towards the quality of this project.

ASSUMPTIONS

This project is a PhD research spanning duration of 3 years. There is no external funding available for this project and thus has to rely on open source and publicly available solutions.

RISKS

It is important to assess and manage the risks that might affect this project. Following table is a summary of potential risks and mitigation plan.

Risk	Severity	Mitigation
Unavailability of experts to verify results	High	Rely on a rich library of Qur'anic science books
Unknown nature of text analysis features	High	Pilot project
Annotation tools unable to handle Arabic	Medium	Build in-house tools
Other research team doing similar work	Low	Find opportunities of collaborative work and create annotation that are not considered yet.

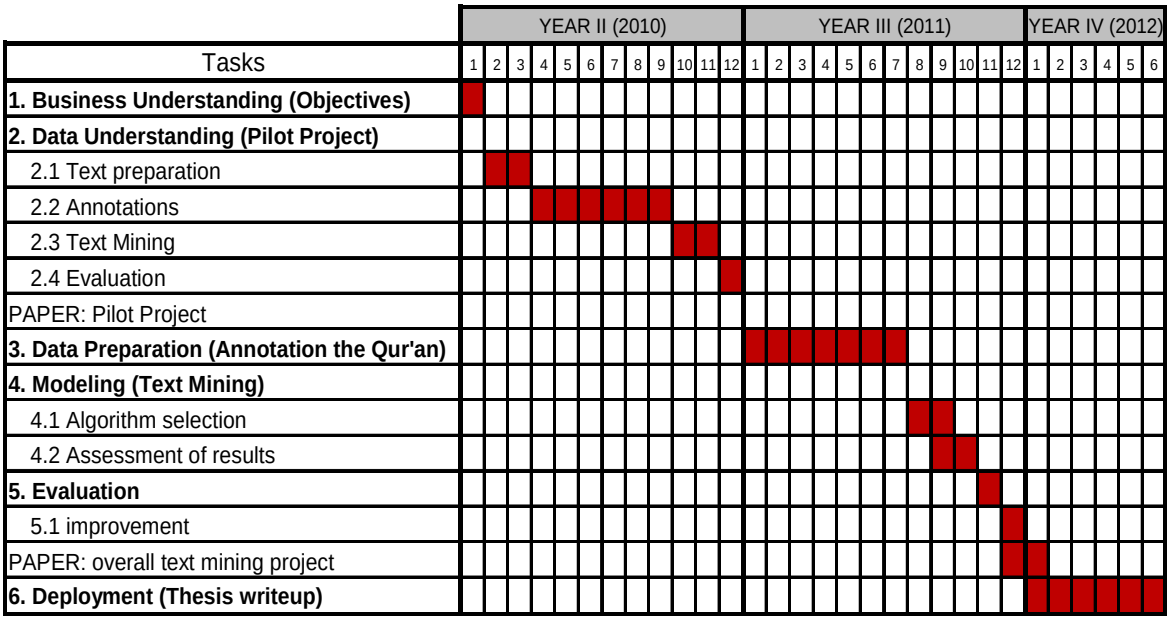
Table 3.4 – Risk assessment

There is a rich library of books on Tafsir and Qur'an science many of them are in the Brotherton Library at Leeds University. Also many books are publicly available on the web in PDF format for download. These resources are likely to contain enough information to validate the result of text mining. Moreover, I have access to address of few Qur'anic scholars in Saudi Arabia who can be consulted.

The project starts with a piloting task on chapters 2 and 7 of the Qur'an. As this project is first of its kind on the Qur'an –to the best of my knowledge-, piloting was attempted to gain some understanding on the type and level of annotation needed.

3.1.3 PROJECT PLAN

We breakdown the effort needed in this project based on industry standards as quoted by CRISP-DM. The project plan has the following distribution of effort: 3% for business understanding; 40% for data understanding (i.e., pilot project); 22% for data preparation (i.e., Qur'an annotation); 10% for modelling; 5% for evaluation; and 20% for deployment (i.e., final thesis write-up). It must be noted that this plan need to be monitored and revised regularly.



3.2 Phase 2- Data Understanding

In this phase we carry on a pilot project with the objective to understand our text (i.e., the Qur'an) more and determine the text mining scope and requirement for the entire Qur'an. Following are the steps within this phase.

3.2.1 Collect the initial data

This involves selecting a representative section of the Qur'an which is likely to reflect the challenges and various issues in the text mining task. We have chosen two chapters from the Qur'an; chapter 2 and chapter 7. This choice was made based on comparing various attributes and features which are likely to influence text mining task. Chapter 2 is a Madani surah (i.e., revealed in Medina) and the largest chapter in the Qur'an and contains a wide range of topics from legislations to stories of prophets. Chapter 7 is a Makki surah and overlaps in some of the topics like the story of Moses and creation of Adam. Together they contain 9,482 words comprising around 12% of the Qur'an. Following table gives some information about these two surahs.

	Chapter 2 – al-Baqarah	**Chapter 7 – al-A'raf**
No. of verses	286	206
No. of words	6141	3341
Where revealed?	Medinah	Makkah
Stories mentioned	Story of creation; Moses; Abraham; Saul and Goliath;	Story of creation; Moses; Noah; Hud; Salih; Shu'ayb; Lot;
Rulings mentioned	Prayer; marriage and divorce; financial transactions; fasting; pilgrimage;	None

Table 3.5 – comparison between chapter 2 and 7

3.2.2 Describe the data

This task involves investigating, analyzing and enriching our sample chapters with annotations as a preparation of text mining step. Through this annotation process, we will gradually build some useful resources for computational analysis of the Qur'an as well.

Multilayered annotation will be carried out in this step including: part-of-speech tagging, syntactic parsing, named entities and their relations, coreference resolution, and semantic role labelling. These layers will be populated by manual and semi-automated process after developing annotation guidelines taking into account Arabic grammar and Tafsir rules.

This extensive annotation process should produce gradually a set of resources which will be necessary for text mining later. These resources include: a database of Qur'anic patterns, a WordNet type of hierarchical relations between Qur'anic words, a Treebank of syntactic parsing of the Qur'an, ontology like relations between Qur'anic concepts, and FrameNet type of lexicon containing semantic frames. These layers will be represented in XML. Figure 3.3 below is a depiction of these layers and resources.

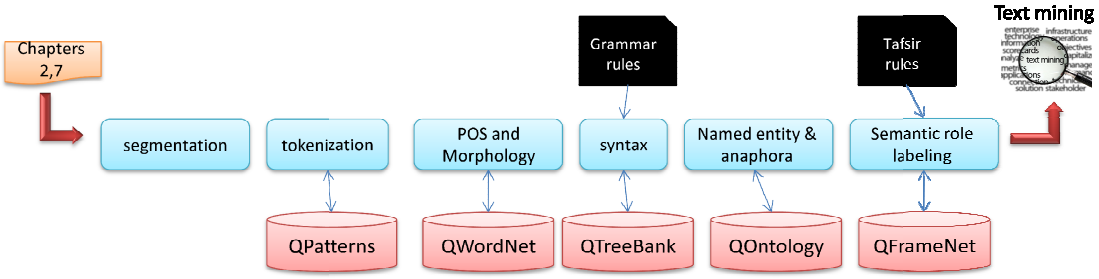


Figure 3.3 – Annotation model for the pilot project

Following is a more elaborate description of each layer. As a case study we chose verse 2:60 to illustrate the presentation of various annotation layers. Figure 3.4 below shows the verse with an English translation.

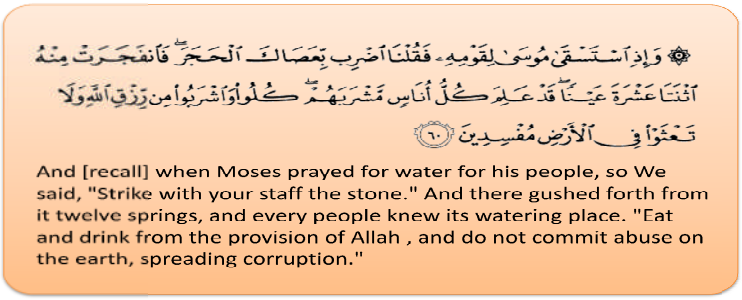


Figure 3.4 – Verse 2:60 as our example verse

TOKENIZATION AND SEGMENTATION

The raw Qur'anic verses are first tokenized into words delimited by whitespace. These words may contain multiple morphemes but that will be broken into single units in the POS layer. Each word of the Qur'an can thus be accessed with a unique ID as shown in figure 3.5 below.

```
<chapter no="2">
...
<verse no="60">
<word no="1">وإذ</word>
<word no="2"> استسقى </word>
<word no="3"> موسى </word>
<word no="4"> لقومه </word>
<word no="5"> فقلنا </word>
<word no="6"> اضرب </word>
<word no="7"> بعصاك </word>
<word no="8"> الحجر </word>
<word no="9"> فانفجرت </word>
<word no="10"> منه </word>
<word no="11"> اثنتا </word>
<word no="12">عشرة </word>
<word no="13">عينا</word>
<word no="14">قد</word>
<word no="15">علم</word>
<word no="16">كل</word>
<word no="16">أناس</word>
<word no="16">مشربهم</word>
<word no="16">كلوا</word>
<word no="16">واشربوا</word>
<word no="16">من</word>
<word no="16">رزق</word>
<word no="16">الله</word>
<word no="16">ولا</word>
<word no="16">تعثوا</word>
<word no="16">في</word>
<word no="16">الأرض</word>
<word no="16">مفسدين</word>
</verse>
...
</chapter>
```

Figure 3.5 – Tokenization of verse 2:60

A single Qur'anic verse often does not correspond to a complete syntactic sentence that conveys a single meaning. Rather, often one verse contains several sentences, and often several verses join to complete one sentence. Qur'anic scholars included within a verse various pause marks to indicate completed meaningful segments within this verse. These marks have become part of the standard Arabic Qur'an today. Figure 3.6 on next page is an example of different pause marks in the verse 2:60 which we considered to be a good candidate of segmentation. While these pause marks are traditionally been used as a guideline for recitation, they play vital role in 'sentencizing' a verse into meaningful units. Annotating the raw Qur'an with these marks is essential in our project as these segments will be the focus when searching for and finding associations and patterns in our mining process. This task is computationally feasible as the Tanzil XML file preserves these marks.

```
<verse no = "60">
<segment no="1">
And [recall] when Moses prayed for water for his
people, so We said, "Strike with your staff the
stone."
</segment>
<segment no ="2">
And there gushed forth from it twelve springs
</segment>
<segment no="3">
every people knew its watering place.
</segment >
<segment no ="4">
Eat and drink from the provision of Allah , and do
not commit abuse on the earth, spreading
corruption.
</segment>
</verse>
```

```
<verse no = "60">
<segment no="1">
وَإِذِ اسْتَسْقَىٰ مُوسَىٰ لِقَوْمِهِ فَقُلْنَا اضْرِب بِّعَصَاكَ الْحَجَرَ ۖ
</segment>
<segment no ="2">
فَانفَجَرَتْ مِنْهُ اثْنَتَا عَشْرَةَ عَيْنًا ۖ
</segment>
<segment no="3">
قَدْ عَلِمَ كُلُّ أُنَاسٍ مَّشْرَبَهُمْ ۖ
</segment >
<segment no ="4">
كُلُوا وَاشْرَبُوا مِن رِّزْقِ اللَّهِ وَلَا تَعْثَوْا فِي الْأَرْضِ مُفْسِدِينَ
</segment>
</verse>
```

Figure 3.6 – Segmenting verse 2:60

QUR'ANIC PATTERNS

Qur'an is characterized by repetitive use of common phrases. Templates using regular expressions can then link such related verses sharing these common patterns. This resource of common patterns will

be very helpful in text mining for related verses and creating clusters of concepts. In the context of our example verse 2:60, comparing it with verses 2:57 and 7:160, we can see many constructs repeating in both stories.

And We shaded you with clouds and sent down on you Al-Manna and the quails, (saying): "Eat of the good lawful things We have provided for you," (but they rebelled). And they did not wrong Us but they wronged themselves. [2:57]	وَظَلَّلْنَا عَلَيْكُمُ الْغَمَامَ وَأَنزَلْنَا عَلَيْكُمُ الْمَنَّ وَالسَّلْوَىٰ ۖ كُلُوا مِن طَيِّبَاتِ مَا رَزَقْنَاكُمْ ۖ وَمَا ظَلَمُونَا وَلَٰكِن كَانُوا أَنفُسَهُمْ يَظْلِمُونَ
And [recall] when Moses prayed for water for his people, so We said, "Strike with your staff the stone." And there gushed forth from it twelve springs, and every people knew its watering place. "Eat and drink from the provision of Allah , and do not commit abuse on the earth, spreading corruption." [2:60]	وَإِذِ اسْتَسْقَىٰ مُوسَىٰ لِقَوْمِهِ فَقُلْنَا اضْرِب بِّعَصَاكَ الْحَجَرَ ۖ فَانفَجَرَتْ مِنْهُ اثْنَتَا عَشْرَةَ عَيْنًا ۖ قَدْ عَلِمَ كُلُّ أُنَاسٍ مَّشْرَبَهُمْ ۖ كُلُوا وَاشْرَبُوا مِن رِّزْقِ اللَّهِ وَلَا تَعْثَوْا فِي الْأَرْضِ مُفْسِدِينَ
And We divided them into twelve descendant tribes [as distinct] nations. And We inspired to Moses when his people implored him for water, "Strike with your staff the stone," and there gushed* forth from it twelve springs. Every people knew its watering place. And We shaded them with clouds and sent down upon them manna and quails, [saying], "Eat from the good things with which We have provided you." And they wronged Us not, but they were [only] wronging themselves. [7:160]	وَقَطَّعْنَاهُمُ اثْنَتَيْ عَشْرَةَ أَسْبَاطًا أُمَمًا ۚ وَأَوْحَيْنَا إِلَىٰ مُوسَىٰ إِذِ اسْتَسْقَاهُ قَوْمُهُ أَنِ اضْرِب بِّعَصَاكَ الْحَجَرَ ۖ فَانبَجَسَتْ مِنْهُ اثْنَتَا عَشْرَةَ عَيْنًا ۖ قَدْ عَلِمَ كُلُّ أُنَاسٍ مَّشْرَبَهُمْ ۚ وَظَلَّلْنَا عَلَيْهِمُ الْغَمَامَ وَأَنزَلْنَا عَلَيْهِمُ الْمَنَّ وَالسَّلْوَىٰ ۖ كُلُوا مِن طَيِّبَاتِ مَا رَزَقْنَاكُمْ ۚ وَمَا ظَلَمُونَا وَلَٰكِن كَانُوا أَنفُسَهُمْ يَظْلِمُونَ

Figure 3.7 – similarity between three verses: compare 2:57 and 2:60 with 7:160. Similar phrases share same colour

MORPHOLOGY AND POS LAYER

This task involves adopting the Crescent corpus (Dukes and Habash 2010) which is publicly available (http://quran.uk.net) and is being manually checked. This corpus was produced using Buckwalter morphology analyzer followed by manual checking. In addition to POS tags, this corpus gives other useful morphological features like gender information, verb forms, plurality and of which person (i.e., first, second or third), type of a verb and grammar case (e.g., nominative, accusative, jussive, etc.). Figure 3.8 is a snapshot from the Crescent corpus depicting morphology analysis of the first four words from verse 2:60.

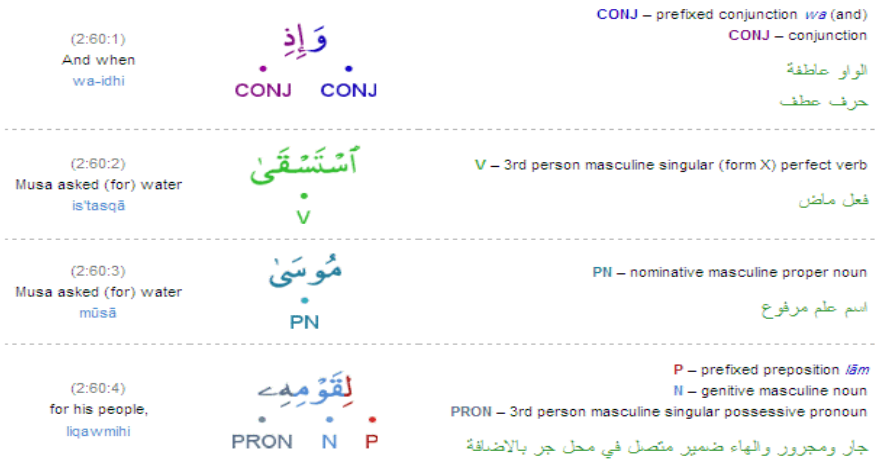


Figure 3.8 – Morphological analysis of part of 2:60 from Crescent corpus

QUR'ANIC WORDNET

Through this task I will organize the Qur'anic vocabulary (verbs, nouns and adjectives) into a hierarchy of semantic relations including: hypernymy-hyponymy (superclass-subclass), synonym-antonym and holonymy-meronemy (part-whole relation). Verbs can exhibit special relations like entailment and cause.

WordNet construction in a new language is usually done by either translating from the English WordNet as was done by (Elkateb et al 2006) for modern standard Arabic, or ground up from scratch as was done by (Piasecki et al 2009) for Polish. I will follow the methodology of the Polish WordNet project and develop semi-automated approach followed by manual verification.

Once finished, this hierarchy of words can be mined for interesting patterns and associations which can give insight into linguistic behaviour of God. Moreover, developing wordnet style synsets help forming conceptual clusters of the Qur'an and can play the role of a lexical ontology. In the course of Qur'anic wordnet construction, books of tafsir and Arabic lexicon can be consulted in cases of ambiguity.

SYNTACTIC ANALYSIS

I will follow traditional Arabic grammar relations in annotating this layer. Standards books are available that exhaustively analyzed the syntax of the entire Qur'an. Figure 3.9 shows syntactic analysis of the first segment of verse 2:60.

```
<verse no = "60">
<segment no="1">
وَإِذِ اسْتَسْقَىٰ مُوسَىٰ لِقَوْمِهِ فَقُلْنَا اضْرِب بِّعَصَاكَ الْحَجَرَ ۖ
</segment>
<syntax>
<range form="إذ"  analysis="اسم مبني على السكون في محل نصب معفول به لفعل محذوف تقديره اذكر" >
<range form="استسقى" analysis="فعل ماض مبني على الفتح">
<range form="موسى" analysis="فاعل مرفوع">
<range form="لقومه" analysis="جار ومجرور">
<range form="فقلنا" analysis="فعل ماض والفاعل والضمير في محل رفع فاعل">
<range form="اضرب" analysis="فعل أمر والفاعل ضمير مستتر تقديره أنت">
<range form="بعصاك" analysis="جار ومجرور">
<range form="الحجر" analysis="مفعول به منصوب بالفتحة">
</syntax>
…
</verse>
```

Figure 3.9 – Syntactic analysis of a segment of 2:60

NAMED ENTITY AND ANAPHORA RESOLUTION

This task involves annotating the Qur'an with named entities based on a developed guideline. This guideline will be constructed following the ENAMEX tagset proposed by the Machine Understanding Conference in 1990 and adding to it domain specific tags. The tagset should be catered for the proper names used in the Qur'an. Majority of Qur'anic proper names are consumed by: names of Allah, names of angels, names of Prophets, their tribes and name of some tribe chiefs who used to reject their prophets. As Qur'an mentions repetitively these names, it is possible to design some patterns for semi-automated extraction and labelling of these names. This kind of annotation will play a major role in text mining tasks when various patterns and associations can be discovered correlating some of these entities. When working with NE recognition task for Arabic, special considerations need to be taken –apart from highly morphological inflections- because capitalization is not used in Arabic for

proper names. Qur'an being a closed system, it is feasible to extract the nouns from POS tagged corpus and annotate the proper names. Figure 3.10 shows a possible representation for verse 2:60.

> And recall when [Moses PROPHET] prayed for water for [his MOSES] people, so [We ALLAH] said, "Strike with [your MOSES] staff the stone." And there gushed forth from [it STONE] twelve springs, and every people knew [its PEOPLE] watering place. "Eat and drink from the provision of Allah , and do not commit abuse on the earth, spreading corruption."

Figure 3.10 – Named entity and anaphora annotation for verse 2:60

QUR'ANIC ONTOLOGY

Finding domain specific entities and defining their relationships will facilitate creation of ontology of the Qur'an. For example, the twelve springs mentioned in verse 2:60 are referring to the twelve tribes of Israel who were the descendents of Jacob referred in verse 2:136. Further following the family tree of prophets would reveal lots of relations. Also, 'eating' in the verse refers to "manna and quails" that was sent down to them by God, as mentioned in verse 2:57. Figure 3.11 shows some of these relations.

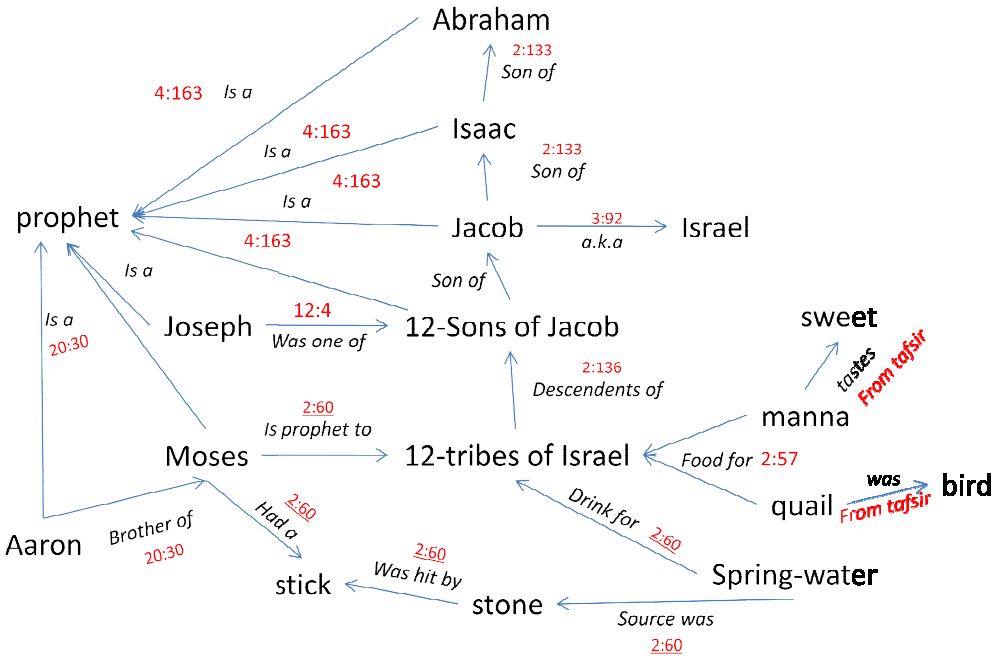


Figure **3.11 – A sample ontology that might evolve from verse 2:60 showing relations between prophets sent to children of Israel with reference to source verses**

SEMANTIC ROLE LABELING

This task adopts the FrameNet semantics (Ruppenhofer et al. 2005) which is an online lexicon of lexical words (mostly verbs) which trigger a prototypical situation called a *semantic frame* and various roles called *lexical elements* participate in completing the overall picture of this frame. Figure 3.12 depicts labeling verse 2:60 with framenet style roles.

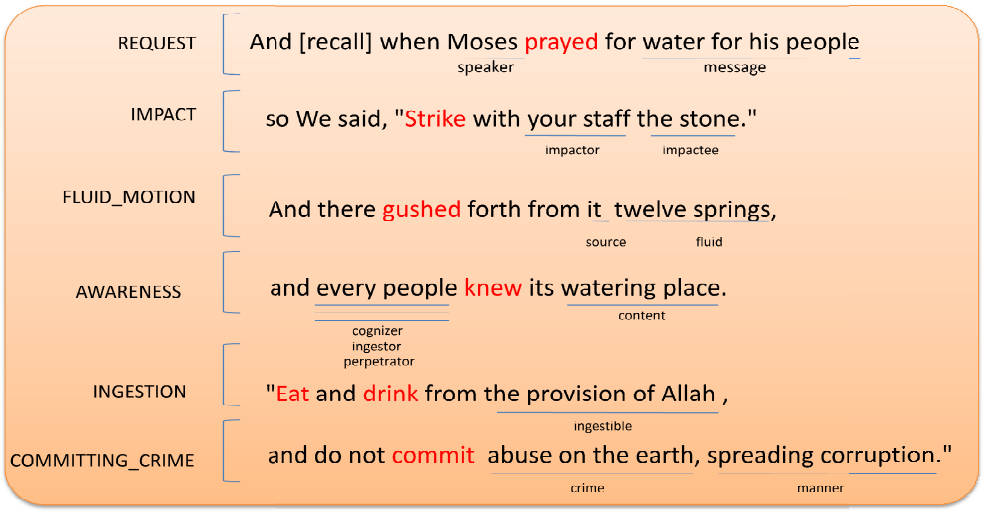


Figure 3.12 – Semantic role labelling of verse 2:60 with frame semantics. Frames are capitalized, Lexical units are in red, and frame elements are underlined.

Prior to annotation process, a comprehensive guide is to be produced. Books of tafsir will be consulted to resolve difficult issues. In the next level, this annotation will be experimented for text mining purpose and may result in needing more or less levels. While annotating at this pilot attempt, techniques will be developed for semi-automating the annotation process for the rest of the Qur'an.

3.2.3 – Explore the Data

This task involves exploring the annotations and resources developed so far covering chapters 2 and 7. Here I will revisit the objectives set earlier and ask question about how the annotation carried so far helps in fulfilling these objectives? I will carry out experiments with existing text mining tools and algorithms on my annotated data. This may reveal the need to include or exclude certain information in annotation. It is expected that there will be so many different text mining usage from the level of annotation described above; I will only choose a subset for this PhD research. This subset is judged by the nature of annotation required for text mining. I want to consider annotating the entire Qur'an only to the level which is feasible within the time allocated in my plan. However, this rich annotation of chapters 2 and 7 along with annotation guidelines will enable annotating the entire Qur'an. In short, this step will define the scope of the annotation process for the entire Qur'an.

3.2.4 – Verify Data Quality

This step analyzes the output of text mining task on my pilot annotation data (i.e., chapters 2 and 7). I will extrapolate from the results and data at hand and think about the possible missing features. What are the main features not covered in chapters 2 and 7 that might affect our text mining task? At this stage, I will be in a position to publish our preliminary findings in a recognized conference.

3.3 – Phase 3 - Data Preparation

This phase continues with the annotation task based on lessons learned and recommendations from the pilot phase. CRISP-DM defines certain tasks in this phase which are very specific to "structured" data stored in a database like select data, clean data, construct data, and integrate data. In the case of text mining, these steps are not needed; rather I will repeat the annotation steps in the pilot project widening our coverage to entire Qur'an but with less layers of annotation based on the pilot study recommendation. Semi-automated techniques will be employed for annotation at this stage. These

techniques will be realized and formalized during the pilot project. Figure 3.13 below shows the annotation steps which look very similar to the pilot annotation steps.

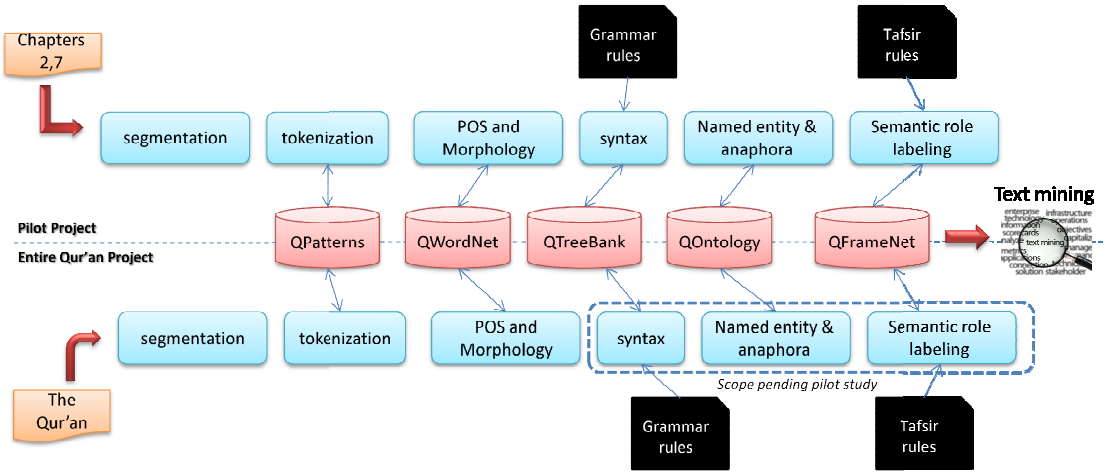


Figure 3.13 – overall picture of the two phase annotation process

3.4 – Phase 4 – Modelling

This is the actual text mining task on the text (i.e., Qur'an) which have been prepared and enriched with annotation in previous phases. In the final step of the pilot project, I should have experimented with few text mining techniques on the initial annotation. Now, based on that experience I will repeat the mining process on the entire Quran. I need to highlight the iterative nature of this task; some text mining experiment may dictate adjusting the annotation to include/exclude certain information. Figure 3.14 is an overview of our text mining model. The specific steps of this phase are described below.

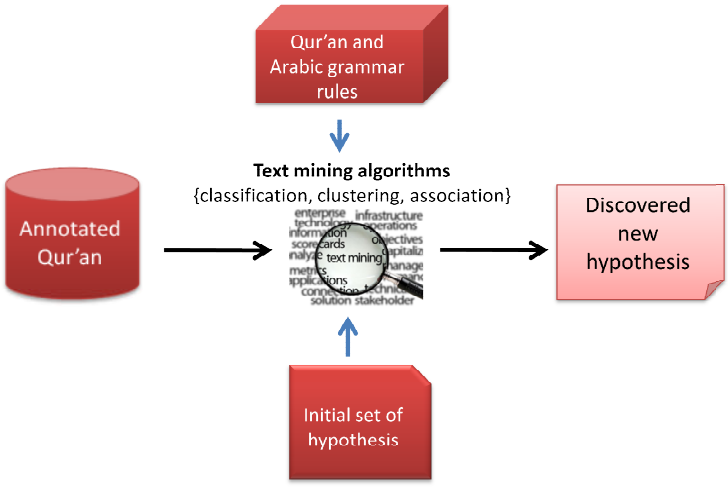


Figure 3.14 – Our Text mining model

3.4.1 – Select the Modelling Technique

Refer to chapter 2 (literature review) for a detailed overview of existing models. The choice of the specific model is influenced by the nature of the new knowledge we are interested in and the nature of the available annotation.

3.4.2 – Generate Test Design

In this task I will speficy the details of the input to the text mining algorithms. This involves defining which data should be trained on, and which one to be tested with. I will also need to define the number of tests and which feature parameters to include for each test, also, how to evaluate the results and check for accuracy.

3.4.3 – Build Model

This involves running the text mining algorithms after setting the appropriate parameters. These parameter setting need to be rationalized and well documented. After each running, the results need to be recorded and documented for later analysis. Also, the outcome of these tests need to be presented in an understandable way.

3.4.4 – Assess the Model

This involves analyzing the outcome of the text mining algorithms for new discoveries and interesting patterns and associations. This step directly links with the objective set at the beginning of the project. Ideally, the results are to be verified by domain experts. As the Qur'an has a rich set of scholarly commentaries available, and given my knowledge of the domain, these results can be verified by me for accuracy. Based on assessment, different parameter settings, feature sets or algorithm can be chosen to build a different model.

3.5 – Phase 5 – Evaluation

This phase involves evaluating our annotated corpus and text mining application against the stated objectives. Does the outcome of our system consider being a new knowledge which could not be discovered otherwise? Do the results satisfy the success criteria set earlier? If not, then how could we improve the system -either by improving the algorithms or annotation- to meet our objectives? It is possible to make minor adjustment to our model and repeat phase 4. As part of evaluation, the system –or part of it- will be made available online for public use and evaluation. It is unlikely at this stage that the project is a complete failure and results are far away from being interesting, as our pilot project would have detected drastic deviation from objectives.

Also at this phase, we document future enhancements and improvements of this project. Being a pioneering text mining project on the Qur'an, this project must have many opportunities for continuation. One of the objectives of the project is to create a framework for other researchers to build on and add more annotation layers and text mining tasks.

At this stage, we have enough experimentation to publish journal paper reporting on our methodology, the annotation process, the framework of text mining and various experiments undertaken.

3.6 – PHASE 6 – DEPLOYMENT

This involves documenting the entire project experience, which is equivalent - in the context of PhD research – to thesis write-up. As a tradition in PhD projects, the last six month will be dedicated for this task. Table 3.5 is a tentative outline of the PhD thesis document.

Chapters	details
Introduction	Description of the topic and motivation
Background	Why choosing the Qur'an
Problem statement	Scope, novelty, originality and contribution
Previous Work	Literature review, available resources
Research Overview	Overall model, CRISP-DM methodology
Pilot Project	Chapters 2,7: annotation and text mining issues
Annotation Process	Annotation guideline for the entire Qur'an
Text Mining	The model, what type of text mining, which algorithms, the feature set
Evaluation and Results	How the results met objectives, validation by scholars
Future Work	Improvement of annotation, features set, algorithms
Conclusion	Contribution of this project for future research on the Qur'an

Table 3.5 – Tentative outline of the PhD Thesis report

CHAPTER 4 – FIRST YEAR PROGRESS

This chapter gives an overview of the progress made so far in my first year at Leeds. The progress is categorized into: training and development progress and those related to my research area.

4.1 – TRAINING AND DEVELOPMENT

In my first session I have enrolled in a number of graduate and undergraduate courses as well as workshops. These sessions proved helpful both in building network and settling in a foreign country as well as in providing background to my research. Table 4.1 below gives a summary.

Date	Course Name	Type/provider
2nd Semester 08/09	Introduction to NLP (Dr. Atwell)	Undergraduate course
2nd Semester 08/09	Language (Dr. Markert)	Graduate course
2nd Semester 08/09	Computer assisted lexicography (Dr. Sheroff)	Undergraduate course
20-Jan-09	Dealing with stress of doing research	SDDU
03-Feb-09	Welcome to Faculty of Engineering	Engineering
17-Feb-09	Finding information for your PhD	Library
26-Feb-09	Safety Course	SDDU
09-Mar-09	Speed PhD	Engineering
26-Mar-09	Effective Research Writing	SDDU
21-May-09	Managing Information for PhD	Library
26-June-09	Network Essentials for Research Students	SDDU
07-Aug-09	Researcher@Leeds	ISS

Table 4.1 – Training Courses

I participated in the training and development needs analysis and compared my progress at month one and month 10 which showed significant improvement. The form is annexed in the appendix.

Also, I participate regularly in weekly discussion group of the NLP team and presented twice my research proposals. In addition I attended few interesting presentations by the "knowledge representation group" (now "intelligence augmented (IA)) and "Arabic and Islamic and Studies".

One of my great networking and training opportunity came when participated in Corpus Linguistics conference in Liverpool on June 2009. Attendance in this conference enables attending many relevant sessions and having informal discussion with many experts in the field. I am attaching 'back to office' report.

4.2 – RESEARCH PROGRESS

I started with developing technical background in terms of investigating available tools and resources for the Qur'an. This involved experimenting with linguistic computing with Unix tools. I obtained the Haifa corpus of Qur'anic POS and morphology and ran experiments with it. I also developed skills with Natural language toolkit (NLTK) under python. Also, I learned data mining techniques with WEKA tool. Chapter 5 gives a detailed account of these experiments.

I spent some time on studying the Arabic Grammar relations in the context of the Qur'an. This resulted in developing a set of relations that can be represented as dependency graphs. This method was then applied to small chapters of the Qur'an. This work is submitted for publication in (Dukes, Eric and Sharaf 2010) for LREC 2010 (attached in Appendix).

I also investigated the possibility of developing a FrameNet type of lexical resource for the Qur'an. This resulted in collecting sample Qur'anic verbs and studying their valences and analyzing the compatibility with FrameNet semantic roles. This work will be continued in the pilot project for chapters 2 and 7. This preliminary investigation resulted in a successful work-in-progress publication in the Corpus Linguistics conferences 2009 in Liverpool where I got opportunity to present the paper and receive feedback. (Paper attached in the appendix).

I also investigated the possible application of a multi-layered annotated Qur'anic corpus for question-answering system. This preliminary study resulted in a submission for an extended abstract for the LREC 2010 conference. The results will be out by February 2010. (Extended abstract attached in appendix).

I have gained much experience in the annotation process of discourse relations in Modern Standard Arabic (MSA) by joining in the annotation project of Al-Saif and Markert. This work gave me insight in developing annotation guidelines for discourse relations in the Qur'an.

In addition to my previous 10 years of exposure to Tafsir books (i.e., scholarly commentary on the Qur'an), I investigated their usage as a background knowledge for the mining process. Especially (As-Sabt 1996; As-Soyouti 1987) which lists a set of rules as well as validation criteria in the Qur'an that can be integrated in our system.

CHAPTER 5 – COMPUTATIONAL EXPERIMENTS

In this chapter, I will report on some computational experiments I did with both raw Qur'an text, and later when a morphologically tagged corpus became available very recently (i.e., the Quranic Arabic Corpus).

5.1 - WORD FREQUENCY DISTRIBUTION

I took an authentic copy of the Arabic Qur'an text from (http://tanzil.info) and tokenized it and did basic counting of words. Here is a list of the 100 most frequent words.

من	2764	ذلك	280	فإن	169	فيه	127
الله	2151	له	275	إذ	165	قد	126
في	1186	الذي	268	عليكم	164	قوم	126
ما	1010	هو	265	والذين	164	عند	119
إن	966	آمنوا	263	الكتاب	163	قبل	118
لا	813	هم	261	والأرض	157	خير	116
الذين	811	وإن	254	فلا	156	لله	116
على	670	قالوا	249	إنا	155	يشاء	116
إلا	664	كل	245	أيها	153	الدنيا	115
ولا	657	فيها	241	منهم	153	إنما	113
وما	642	والله	241	عذاب	150	ولكن	112
أن	638	كانوا	229	بعد	149	ربهم	111
قال	416	عن	223	إنه	147	مما	111
إلى	405	إذا	221	عليه	146	ولو	111
لهم	373	ربك	220	حتى	142	الحق	109
يا	350	يوم	217	بالله	139	السماء	109
ومن	342	عليهم	214	وهم	137	منكم	107
ثم	340	شيء	190	أولئك	133	ربنا	106
لكم	337	كفروا	189	وإذا	132	عليم	106
به	327	كنتم	188	أم	131	النار	102
كان	323	هذا	188	إني	131	ربكم	102
قل	294	السماوات	182	رب	130	فلما	101
بما	292	الناس	182	موسى	129	ألا	99
الأرض	287	لم	181	ولقد	129	أنزل	95
أو	284	وهو	171	بل	127	ربي	94

Table 5.1 – Frequency distribution of the first 100 words

This list contains lots of function words like pronouns, particles and prepositions. Following is a list of 150 most frequent function words (Table 5.2 next page).

39	وعلى	55	أنت	78	ألم	131	إني	261	هم	2764	من
38	كلا	54	لقد	78	أنتم	129	ولقد	254	وإن	1186	في
38	وكذلك	51	كنت	78	بها	127	بل	241	فيها	1010	ما
37	بكم	50	إنك	77	إليك	127	فيه	223	عن	966	إن
37	عنه	48	علينا	76	إليه	126	قد	221	إذا	813	لا
37	لولا	47	عما	75	عنهم	119	عند	214	عليهم	811	الذين
35	فبأي	47	ليس	73	لما	118	قبل	188	كنتم	670	على
35	منا	47	هي	72	وإذ	113	إنما	188	هذا	664	إلا
35	وكانوا	47	ولم	71	لك	112	ولكن	171	وهو	657	ولا
35	يكون	46	فأولئك	69	غير	111	مما	169	فإن	642	وما
34	بكل	46	هذه	68	وأن	111	ولو	165	إذ	638	أن
34	معه	45	أفلا	67	هل	107	منكم	164	عليكم	405	إلى
33	أولم	45	عليها	66	أنا	102	ربكم	164	والذين	373	لهم
33	قبلك	45	وأنتم	65	نحن	101	فلما	156	فلا	350	يا
33	وفي	44	ذلكم	63	إنهم	99	ألا	155	إنا	342	ومن
33	ولئن	43	وقد	63	كنا	94	لمن	153	أيها	340	ثم
32	إليكم	43	ولهم	63	لي	90	فمن	153	منهم	337	لكم
32	وإنا	41	أنهم	62	كيف	88	منه	149	بعد	327	به
32	ولما	41	بغير	59	لن	87	فإذا	147	إنه	323	كان
32	يكن	41	بهم	58	بينهم	86	فما	146	عليه	292	بما
31	إنكم	41	فهم	58	كما	86	كذلك	142	حتى	284	أو
31	بينهما	41	لعلهم	58	مع	86	منها	137	وهم	280	ذلك
31	ممن	41	لها	57	التي	84	وكان	133	أولئك	275	له
30	فإنما	40	إليهم	57	عليك	83	لنا	132	وإذا	268	الذي
29	بينكم	40	هؤلاء	57	فقد	79	لو	131	أم	265	هو

Table 5.2 – Most frequent 150 function words in the Qur'an

Taking out the function words brings the list of content words for the Quran, as shown in table 5.3 next page. It is evident from the list the inherent problem of Arabic language where many prefix and suffixes are attached to the word without any whitespace. For example, the word 'Allah' alone counts 2151, but when we add those tokens with clitics like 'waAllah' (and/by Allah), 'biAllah' (by Allah), 'liAllah' (for Allah) we have total of 2,657 occurrence. Also, the word 'he said' *qaAla* appears in our list as different from the imperative 'say' *qulo.* Taking all 'say' roots and adding them brings the frequency to 1,105 occurrences.

This problem makes it a supportive case for considering annotated corpus where morphological tagging is added with information about lemma and root.

Allah	الله	2151	their lord	ربهم	111	world hereafter	الأخرة	71	he came	جاء	57
he said	قال	416	the truth	الحق	109	he knows	يعلم	71	the garden/jannat	الجنة	56
say	قل	294	the sky	السماء	109	day of resurrection	القيامة	70	people of	أهل	56
the earth	الأرض	287	our lord	ربنا	106	Pharaoh	فرعون	67	you know	تعلمون	56
they believed	آمنوا	263	all knowing	عليم	106	and his messenger	ورسوله	66	he wishes	شاء	56
they said	قالوا	249	the fire	النار	102	their hearts	قلوبهم	65	they know	يعملون	56
all/eat	كل	245	your lord	ربكم	102	on that day	يومئذ	65	did he knew?	أعلم	55
and/by Allah	والله	241	sent down	أنزل	95	the oppressor	الظالمين	64	a little	قليلا	55
your lord	ربك	220	my lord	ربي	94	the exalted in might	العزيز	64	and they did	وعملوا	53
day	يوم	217	below	دون	91	the life	الحياة	63	the infidels	الكافرين	52
some/any thing	شيء	190	path/way	سبيل	87	the satan	الشيطان	62	signs	آيات	52
disbelieve	كفروا	189	they believe	يؤمنون	86	the people/tribe	القوم	62	painful	أليم	52
the heavens	السماوات	182	the punishment	العذاب	85	people of	أصحاب	62	messenger	رسول	52
people	الناس	182	and he said	وقال	85	Abraham	إبراهيم	62	forgiveful	غفور	52
the Book	الكتاب	163	clear	مبين	84	the gardens	جنات	62	sone of	بني	51
and the earth	والأرض	157	you know	تعملون	83	good deeds/women	الصالحات	61	for the people	لقوم	51
punishment	عذاب	150	he created	خلق	83	the worlds	العالمين	61	for mankind	للناس	51
by Allah	بالله	139	they know	يعلمون	82	some	بعض	61	they say	يقولون	51
lord	رب	130	the believer	المؤمنين	78	merciful	رحيم	61	we have sent	أرسلنا	50
Moses	موسى	129	something	شيئا	77	and they said	وقالوا	61	nation	أمة	49
people/tribe	قوم	126	with truth	بالحق	74	he knew	علم	60	all	جميعا	49
good	خير	116	god	إله	73	the night	الليل	59	severe	عظيم	49
for Allah	لله	116	his lord	ربه	73	the mankind	الإنسان	58	runs (fem)	تجري	48
he wished	يشاء	116	their selves	أنفسهم	72	the prayer	الصلاة	58	the religion	الدين	47
this world	الدنيا	115	hellfire	جهنم	72	with our signs	بآياتنا	57	sign	آية	47

Table 5.3 – The 100 most frequent content words in the Qur'an

The above result was based purely on counting the words without any relative measure of likelihood against a reference corpora. I could not find any reference corpora that captures classical Arabic, so I made a comparison with the Corpus of Contemporary Arabic (CCA) developed by Latifa As-Sulaiti at Leeds and publicly available at http://www.comp.leeds.ac.uk/eric/latifa/research.htm. It contains 587707 words from various genres like politics, fiction, religion, health, sports, education, economics and science. Following table gives the results.

rank	Word	Meaning	CCA	Qur'an	LL-score
1	الله	Allah	535	2151	2452.62
2	الذين	those	603	811	697.53
3	إن	sure	1160	966	656.87
4	وما	and those	372	642	603.18
5	إلا	except	726	664	476.71
6	ولا	and do not	741	657	463.52
7	لكم	yours	26	337	433.48
8	قل	say	9	294	393.71
9	لهم	theirs	175	373	372.7
10	آمنوا	believe	2	263	360.78
11	قال	say	306	416	359.31
12	قالوا	they said	15	249	324.76
13	والله	by Allah	22	241	306.55
14	ربك	your lord	2	220	301.26
15	ما	no	2921	1010	288.22
16	يا	O (vocative)	334	350	269.75
17	هم	they	113	261	266.26
18	كنتم	you were	4	188	254.08
19	السماوات	the skies	5	182	244.46
20	عليهم	on them	70	214	232.4
21	وإن	and if	161	254	231.67
22	بما	for	272	292	227.7
23	الأرض	the earth	305	287	209.28
24	عليكم	on you	15	164	208.58
25	إنا	surely we	4	155	208.53

rank	Word	Meaning	CCA	Qur'an	LL-score
26	والذين	and those	17	164	206.64
27	عذاب	punishment	2	150	204.42
28	والأرض	and earth	12	157	202.06
29	كانوا	they were	160	229	201.71
30	أيها	o you who	35	153	176.71
31	رب	lord	8	130	169.37
32	به	with it	572	327	168.98
33	موسى	Moses	9	129	166.93
34	إني	me	12	131	166.59
35	بالله	by Allah	25	139	165.85
36	قوم	tribe	7	126	164.99
37	يشاء	will	3	116	156.06
38	ومن	and from	699	342	152.35
39	يوم	day	258	217	148.44
40	لا	no	2983	813	147.75
41	أولئك	they	39	133	147.46
42	منكم	from you	2	107	144.99
43	عليم	knowlegeable	2	106	143.61
44	ربنا	our lord	4	106	141.01
45	ربكم	your lord	1	102	139.56
46	أنزل	sent down	1	95	129.88
47	ولقد	and certainly	59	129	129.77
48	لله	for Allah	35	116	127.94
49	خير	good	36	116	127.27
50	وهم	and they	86	137	125.37

Table 5.4 – Frequency distribution of the Qur'an with Loglikelihood score against the Contemporary Corpus of Arabic (CCA)

This list does not bring in new words which were absent in the earlier frequency list, but reorders the ranks of some of the words (e.g., the word Moses was advanced to rank 11 -considering only content words- after being in rank 20 without LL estimates). As we lack a reference corpus for classical Arabic cotemporary to the time of the Qur'an revelation, and as the results without LL show similar results, I will continue the analysis without considering LL comparison with CCA.

5.2 - N-GRAM

Next, I continue my analysis by counting the most frequent n-grams in the Qur'an starting from bigrams until 5-gram. I used basic Unix commands of `tail`, `paste`, `sort` and `uniq` as follows.

```
# First take a raw file each token in a line and create another file
#translating by one
-bash-3.2$ tail -n+2 qtokens.dat>qtokens.next
#repeat the proces to the desired level
-bash-3.2$ tail -n+2 qtokens.next>qtokens.next2
-bash-3.2$ tail -n+2 qtokens.next2>qtokens.next3
-bash-3.2$ tail -n+2 qtokens.next3>qtokens.next4
#now align all these files side-by-side, sort then, extract the unique with
#frequencey
-bash-3.2$ paste qtokens.dat qtokens.next qtokens.next2 qtokens.next3
qtokens.next4|sort |uniq -c>5-gram.txt
```

Following tables show the results.

205	إن الله	indeed Allah is
184	الذين آمنوا	those who believed
176	في الأرض	on the earth
143	يا أيها	O you who
134	الذين كفروا	those who disbelieve
133	السماوات والأرض	the heavens and the sky
92	أيها الذين	those who
92	من قبل	from before
89	من الله	from Allah
87	كل شيء	every thing
84	إن الذين	indeed those
84	من بعد	from after
83	من دون	from without
80	إن كنتم	if your were
72	دون الله	without allah
72	على الله	on Allah
71	الله من	Allah who
71	في السماوات	in heavens
69	سبيل الله	path of Allah
69	ما في	which in

Table 5.5 – Twenty most frequent bigrams in the Qur'an with frequencies

92	الذين	أيها	يا	o you who
89	آمنوا	الذين	أيها	you who believe
71	الله	دون	من	other than Allah
52	شيء	كل	على	over every thing
50	الصالحات	وعملوا	آمنوا	believe and deed good
50	ذلك	في	إن	in it is
44	الله	سبيل	في	in path of allah
39	السماوات	في	ما	whatever in the heavens
37	إلا	إله	لا	no god except
36	وعملوا	آمنوا	الذين	those who believe and deed
34	تحتها	من	تجري	underneath runs
34	الأنهار	تحتها	من	from underneath river
33	الله	إن	الله	allah indeed allah
33	قدير	شيء	كل	everything able
31	تكذبان	ربكما	آلاء	bounty of your lord deny
31	ربكما	آلاء	فبأي	on which bounty of your lord
30	هو	إلا	إله	god except he
30	والأرض	السماوات	في	on the heavens and earth
29	لا	الله	إن	indeed allah no
28	في	وما	السماوات	the heaven and what in

Table 5.6 – Twenty most frequent tri-grams in the Qur'an, highlighted two meanings phrases

89	ياأيهاالذينآمنوا	o you who believed
36	الذينآمنواوعملواالصالحات	those who believed and deed good
34	تجريمنتحتهاالأنهار	underneath runs river
33	علىكلشيءقدير	over all things competent
31	فبأيآلاءربكماتكذبان	so which of the favors of your lord would you deny?
30	لاإلهإلاهو	no god except him
28	السماواتومافيالأرض	heavens and whatever on the earth
28	فيالسماواتومافي	in heavens and whatever on
28	مافيالسماواتوما	whatever in heavens and whatever
27	أيهاالذينآمنوالا	those who believed do not
26	جناتتجريمنتحتها	gardens underneath which runs
24	إنفيذلكلآيات	in it indeed signs
20	إنفيذلكلآية	indeed in it a sign
20	تحتهاالأنهارخالدينفيها	underneath it river stay forever in it
19	منتحتهاالأنهارخالدين	from underneath it rivers stay forever
17	السماواتوالأرضومابينهما	the heavens and the earth and whatever in between
17	اللهعلىكلشيء	allah over every thing
17	ولكنأكثرالناسلا	but most of the people do not
14	فيذلكلآياتلقوم	in it signs for the people
12	الذيخلقالسماواتوالأرض	who created the heavens and the earth

Table 5.7 – Twenty most frequent 4-grams in the Qur'an, highlighted are the meaningful phrases

28	في السماوات وما في الأرض	in the heavens and whatever in the earth
28	ما في السماوات وما في	whatever in the heavens and whatever on
27	يا أيها الذين آمنوا لا	o you who believe do not
26	جنات تجري من تحتها الأنهار	garderns underneath which runs rivers
19	تجري من تحتها الأنهار خالدين	rivers runs underneath stay forever
19	من تحتها الأنهار خالدين فيها	from underneath rivers stay forever therein
14	إن في ذلك لآيات لقوم	indeed in these signs for the people
12	خوف عليهم ولا هم يحزنون	fear for them nor shall they grieve
11	الله على كل شيء قدير	Allah over everying competent
11	ولكن أكثر الناس لا يعلمون	but most of the people do not know
11	يا أيها الذين آمنوا إذا	o you who believe if
10	أصحاب النار هم فيها خالدون	people of fire stay therein forever
10	إن الذين آمنوا وعملوا الصالحات	indeed those who believe and do good deeds
9	أظلم ممن افترى على الله	unjust than he who forges against Allah
9	اعبدوا الله ما لكم من	worship Allah you have no from
9	الله ما لكم من إله	Allah you have no god
9	إن الله على كل شيء	indeed allah over everything
9	له ما في السماوات وما	to him belongs what ever in the Heavens and whatever
9	ما لكم من إله غيره	you have no god except Him
9	والله على كل شيء قدير	and Allah over everything competent

Table 5.8 – Twenty most frequent 5-grams in the Qur'an

5.3 - Makki and Madani Chapters

One of the topics that Qur'anic scholars always discussed is the Makki and Madani chapters of the Qur'an. The 114 chapters are classified by scholars as either Makki or Madani based on their chronological order of revelation in reference with the migration of Prophet Muhammad from Makkah to Medinah. In respect with this classification, not all chapters are agreed between scholars, and for a particular Makki chapter, there could be few verses revealed afterwards in Medinah, or vice versa. I will investigate this issue further in the next chapter when learning classifiers in Weka.

Next, I created two files of the raw Qur'an, one which contains only makki chapters, and another only madani chapters. Makki chapters contained 47,643 words (61.2%) of them 6,358 hapax legomena (13.3%) and Madani chapters 30,161 words (38.8%) of them 4,621 hapax legomena (15.3%). Here are the first 100 most frequent words among Makki chapters after removing the function words.

807	الله	Allah
339	قال	he said
200	الأرض	the earth
200	قل	say
191	ربك	your lord
180	قالوا	they said
168	يوم	day
121	السماوات	the heavens
112	والأرض	and the earth
106	موسى	Moses
96	قوم	people
93	عذاب	punishment
92	الناس	mankind
114	رب	lord
89	ربي	my lord
86	السماء	the heaven
79	ربهم	their lord
78	كفروا	disbelieved
77	آمنوا	believed
77	ربكم	your lord
75	مبين	clear
75	وقال	and said
70	لله	for Allah
69	الكتاب	the book
69	ربنا	our lord
68	خلق	created
65	العذاب	the punishment
65	يؤمنون	they believe
64	الدنيا	this world
62	الحق	the truth
61	خير	good
60	بالحق	with truth
60	يعلمون	they know
59	فرعون	Pharaoh
59	يشاء	he wills
58	النار	the fire
55	ربه	his lord
53	إله	god
50	الإنسان	the man
50	العالمين	the worlds
49	العزيز	the exalted in Mighty
49	بآياتنا	with our signs
49	جاء	came
48	القيامة	resurrection
47	الآخرة	hereafter
47	وقالوا	and they said
46	الليل	the night
45	الظالمين	the oppressors
45	جهنم	the hell
44	أرسلنا	we sent
44	الحياة	the life
44	شاء	wills
43	لقوم	for people
42	أعلم	knows better
42	الرحمن	the merciful
42	علم	he knew
41	أصحاب	the people of
40	الجنة	the paradise
40	ويوم	and day
39	الصالحات	good
38	تعملون	you know
38	شيئا	something
38	كذبوا	belie
37	أنزل	sent down
37	أنفسهم	theirselves
37	يعملون	they know
36	القرآن	the Quran
36	آية	sign
36	قومه	his people
35	أمة	nation
35	بالله	by Allah
35	تعلمون	you know
35	عليم	all knower
34	القوم	the people
34	آيات	signs
34	والله	by Allah
33	اليوم	today
33	بني	erect/children
33	جاءهم	came to them
33	ظلموا	did unjust
33	كتاب	book
33	وجعلنا	and we made
32	الشيطان	the satan
32	جعل	made
32	جنات	gardens
32	نفس	soul
32	وعملوا	and did
32	يعلم	he knows
31	رحمة	mercy
31	يقولون	they say
30	إبراهيم	Abraham
30	صالحا	good person
30	قليلا	little
29	الأولين	the ancients
29	الدين	religion
29	الساعة	the Hour
29	آياتنا	our signs
28	ذكر	male/reminder
28	قوما	a tribe
28	نوح	Noah

Table 5.9 – The 100 most frequent content words [Makki]

1344	الله	Allah
207	والله	by/and Allah
186	آمنوا	believed
111	كفروا	disbelieved
104	بالله	by Allah
94	الكتاب	the Book
94	قل	say
90	الناس	people
87	الأرض	the Earth
77	شيء	something
77	قال	he said
71	عليم	all knowing
69	قالوا	they said
64	ورسوله	and his messenger
63	سبيل	path/way
61	السماوات	the heavens
58	أنزل	he sent down
57	عذاب	punishment
57	يشاء	he wills
55	خير	good
53	المؤمنين	the believers
51	الدنيا	this world
50	قلوبهم	their heart
49	يوم	day
47	الحق	the truth
46	لله	for allah
45	تعملون	you know
45	والأرض	and the earth
44	النار	the fire
42	رحيم	merciful
42	قبل	before
40	غفور	forgiver
39	شيئا	something
39	يعلم	he knows
37	الرسول	the messenger
37	الصلاة	prayer
37	أهل	people of
37	ربنا	our lord
37	واتقوا	and fear
35	أنفسهم	theirselves
32	إبراهيم	Abraham
32	ربهم	their lord
31	آلاء	favors
31	تكذبان	be lie
31	ربكما	your lord (dual)
30	الشيطان	the satan
30	تجري	runs
30	جنات	gardens
30	قوم	people
30	يحب	loves
29	الأنهار	rivers
29	ربك	your lord
29	كثيرا	many
28	القوم	the tribe
28	النبي	the prophet
28	تحتها	underneath
28	للناس	for people
28	مريم	Mary
27	الكافرين	the infidels
27	جميعا	all of them
27	جهنم	the Hell
27	حكيم	wise
26	الآخر	the other
26	إليك	
26	خالدين	forever
26	خيرا	good
25	أليم	painful
25	جناح	blame
25	ربكم	your lord
25	رسول	messenger
25	قليلا	little
24	الآخرة	hereafter
24	أنفسكم	yourselves
24	قدير	competent
24	يهدي	guides
23	أبدا	always
23	ابن	son of/build
23	السماء	the heaven
23	أوتوا	they got
23	عظيم	huge
23	موسى	Moses
22	الحرام	sacred
22	الصالحات	good
22	القيامة	resurrection day
22	النساء	women
22	مثل	like
22	واليوم	and the day
22	يريد	wants
22	يعلمون	they know
21	أصحاب	people of
21	الموت	the death
21	تعلمون	you know
21	عذابا	a punishment
21	عظيما	huge
21	عليما	all knowing
21	فضل	favor
21	وعملوا	and they deed
21	يؤمنون	they believe
20	الزكاة	Zakat
20	العذاب	the punishment

Table 5.10 – The 100 most frequent content words [Madani]

Finally, here are few observations from the above two frequency lists:

1. The word ‘Allah’ tops the list in all categories emphasizing the central point of Islam surrounding the monotheism creed. This is also true for the word ‘rabb (lord)’.
2. The verb ‘say’ appears frequently (ranked 2nd in Makki and 7th in Madani) where either Allah ‘says’ to mankind through this Qur’an, or commands is made by Allah to Muhammad to ‘say’ some message to the people.
3. Makki chapters refers more to signs of Allah in ‘nature’ like the Earth, skies, etc. to establish logical proof of the sovereignty of Allah over the nature as opposed to the contemporary belief of the polytheist Makkan people.
4. Makki surah does emphasize on ‘warning’ referring to words like ‘punishment’, ‘hell’, etc.
5. Stories of prophets like ‘Moses’ appears more in Makki chapters.
6. Madani chapters mentions ‘zakat’ which was obligated in Medina period.

QURANIC ARABIC CORPUS (QAC)

Very recently a POS tagged corpus was released on [http://quran.uk.net]. I repeated the process for Makki and Madani considering this corpus. As this corpus indicates lemmas and root words, the results are more accurate than working with raw text. From QAC I used the following tag-set as function words and excluded in content word analysis.

TAG	POS
'T'	Time adverb
'DEM'	demonstrative pronoun
'NEG'	negation
'REL'	relative pronoun
'P'	particle
'PRON'	pronoun
'ACC'	accusative particle
'CONJ'	conjunction
'RES'	restriction particle
'COND'	conditional particle
'INTG'	interrogative
'SUB'	subordinating conjunction
'INC'	inceptive particle
'LOC'	location adverb
'CERT'	certainty particle
'ANS'	answer particle
'SUP'	surprise particle
'RET'	retraction particle

Table 5.11 – QAC Tags used to exclude function words from my analysis

Following table gives a comparative analysis of the frequency of content keywords in all chapters against Makki and Madani Chapters. From this analysis we could note the following points.

1. Makki chapters stress on the word ‘say’ where Muhammad is repeatedly commanded to say the message of Islam to the people of Makkah. Makki chapters mentions lots of stories of previous prophets and their conversation with their people.
2. as a result of the above the word ‘people’ also appears a lot in makki chapters. This is in the context of a dialogue between a prophet –including Muhammad – and his tribe or people.
3. adding POS annotation in our analysis improved the accuracy of results, for example, in our previous experiment for Makki chapters we got 807 occurrence of the word ‘Allah’ followed by 339 of the word ‘say’, but here we recalled 2472 instance of ‘say’ and 1884 occurrence of the word ‘Allah’ this brings both precision and recall higher.

	All			Makki			Madani		
Rank	Freq.	word	meaning	Freq	word	meaning	Freq	word	meaning
1	2628	ٱللَّه	Allah	2472	قول	say	3374	ٱللَّه	Allah
2	1703	قول	say	1884	ٱللَّه	Allah	878	امن	believe
3	1385	كون	be	1812	كون	be	934	قول	say
4	947	ربب	lord	1476	رَبِّ	lord	918	كون	be
5	842	رَبِّ	lord	1476	ربب	lord	698	علم	know
6	858	امن	believe	956	علم	know	570	كفر	disbelieve
7	827	علم	know	900	قوم	people	466	رسل	messenger
8	850	قوم	people	738	امن	believe	422	رَسُول	messenger
9	543	اتي	come	670	اتي	come	418	ربب	lord
10	512	كفر	disbelieve	640	ارض	earth	416	اتي	come
11	507	رسل	messenger	640	أَرْض	earth	408	رَبِّ	lord
12	501	شيا	something	606	شيا	something	400	قوم	people
13	450	ارض	Earth	576	قَوْم	people	398	شيا	something
14	450	أَرْض	Earth	548	رسل	messenger	318	كتب	book
15	383	يَوْم	day	546	يَوْم	day	282	وقي	fear
16	383	يوم	day	546	يوم	day	278	عذب	punish
17	377	قَوْم	people	532	جعل	make	278	انس	man
18	370	عذب	punish	524	سمو	sky	270	عمل	do
19	367	سمو	sky	472	راي	see	266	قتل	kill
20	348	عمل	do	462	عذب	punish	264	نَّفْس	soul
21	348	كلل	everyting	458	كلل	everything	264	نفس	soul
22	346	كُلّ	everyting	456	كُلّ	everything	262	غفر	forgive
23	343	جعل	make	454	كفر	disbelieve	262	إِنسَٰن	mankind
24	334	انس	man	440	عبد	slave	260	ارض	earth
25	328	رَسُول	messenger	438	ءَايَٰت	signs	260	أَرْض	earth
26	327	رحم	mercy	438	رحم	mercy	254	نزل	come down
27	320	عَذَاب	punish	438	سَمَآء	sky	248	ولي	friend
28	319	راي	see	430	خلق	create	248	مُؤْمِن	believer
29	314	كتب	book	426	عَذَاب	punish	238	كِتَٰب	bok
30	313	هدي	guidance	426	عمل	do	238	كلل	everything

Table 5.12 – The 30 most frequent words in the Qur'an and in Makki or Madani Chapters

Then, I wanted to see those lemmas that appear exclusively in makki and madani chapters. This feature will help me later to define feature set for machine learning algorithms.

Here are the results.

	Only Makki			Only Madani		
Rank	Freq.	word	meaning	Freq	word	meaning
1	60	كَلَّا	never	50	جُنَاح	sin
2	54	يُوسُف	Joseph	30	نَصْرَانِيّ	a christian
3	44	قَرْن	stay inside	30	نصاري	christian
4	44	سِحْر	witch craft	28	حُدُود	borders
5	40	فطر	innovate	26	قِتَال	fighting
6	40	كَشْف	clear	24	رِضْوَٰن	satisfaction
7	38	أَخَا	brother	22	مَسِيح	Jesus
8	32	فاد	ransom	22	ثَوَّاب	reward
9	32	فُؤَاد	heart	20	أَعْرَاب	bedowins
10	32	كِيل	weigh	18	حَجّ	Hajj
11	28	أمين	trustworthy	18	مَّرِيض	sick
13	26	ضُرّ	hardship	18	خَبِيث	bad
14	26	قَادِر	competent	18	يَهُودِيّ	a jew
15	26	وصف	attribute	18	يهود	Jews
16	24	جحد	deny	16	برّ	piety
17	24	مَّوْعِد	apointment	16	مُحْصَنَٰت	chaste
18	24	شمل	left	16	صَدَقَٰت	charity
19	24	نطق	utter	16	وَصِيَّة	will
20	24	سِنِين	year	14	أَدَّى	perform
21	24	سجن	glory	14	هَدْي	guidance
22	24	سلك	went it	14	رِّبَوٰا	increase
23	22	مَّعْلُوم	known	12	بُهْتَٰن	wrong
24	22	مَجْنُون	mad	12	فَرِيضَة	obligation
25	22	شَقَو	go astray	12	أُمِّيِّين	wish
26	22	شُعَيْب	Shoaib (prophet)	12	حرف	edge
27	20	شَفِيع	interceed	12	مُّؤْمِنَة	she believer
28	20	بطش	assult	12	ميل	incline
29	20	أَبَا	father	12	سَعَة	wide
30	20	حيق	surround	12	سَكِينَة	peace

Table 5.13 – Words appearing exclusively in either Makki or Madani chapters

Here are some observations from this list:

1. Some prophets like Joseph are only mentioned in Makki chapters
2. Prophet Muhammad established a state and started to contact other world religions only after he migrated to Medinah, this is why words like 'Christian', 'Jew' appears only in Madani chapters.
3. Acts of worship like 'Hajj' is obligated lately and hence is only mentioned in Madani chapters.
4. Muhammad only started military fighting after migration to Medinah, and hence the word 'fighting' was mentioned 26 times only in Madani chapters.
5. Many Madani chapters detail on Islamic rulings on marriage and family life, and hence many reference to women like 'chaste', 'believing women' are mentioned.
6. Some of the strong warning and threatening words can be see in Makkah chapters only like 'assault', 'never', 'surrounded', etc.

5.4 - Verse Similarity Application

As highlighted several times in this report, one of the interesting feature in the Qur'an is the presence of similar but scattered verses in the Qur'an. It is one of the objective of this research is to desgin text mining applications to capture all these similar verses through mulit-layered annotation of the Qur'an. In this experiment, I created a python program to input a user verse and print all other verses that share 1/3 of the total number of content keywords of the source verse. Following is a snapshot of the program.

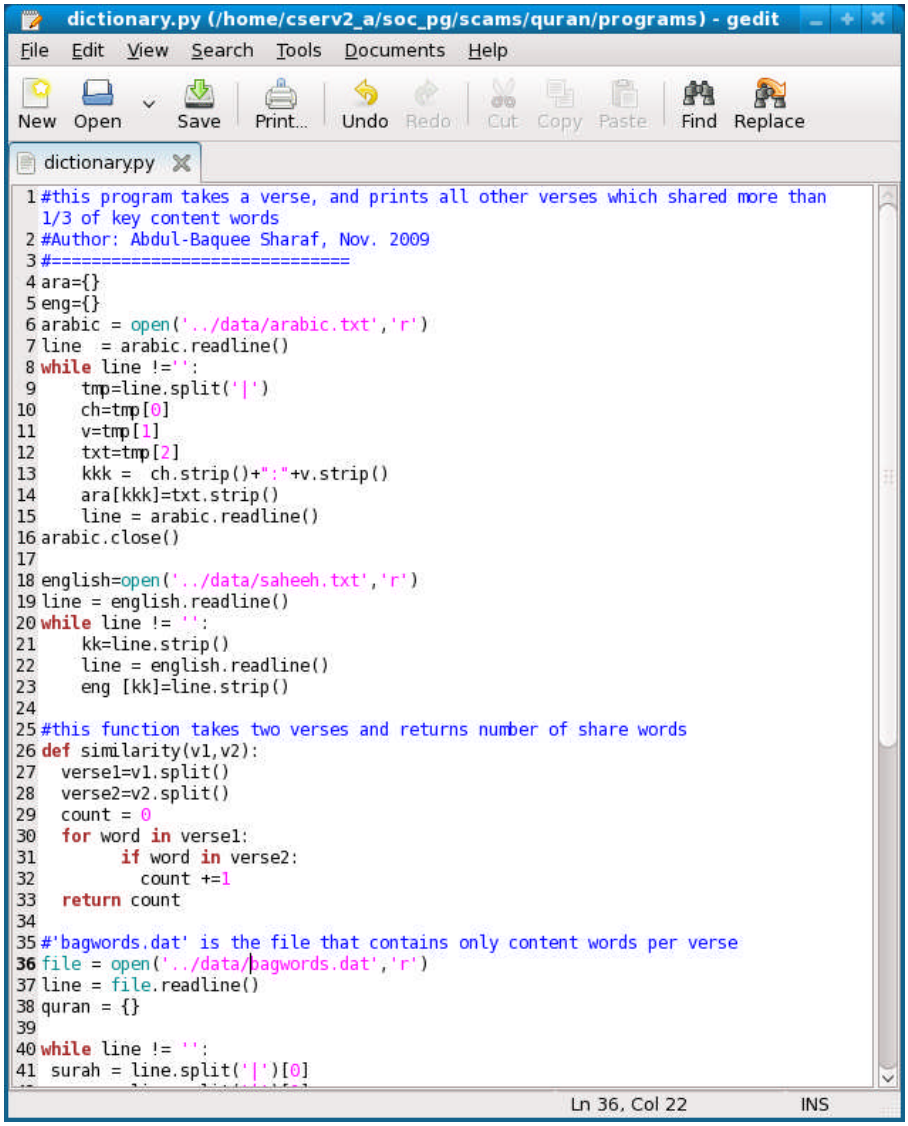


Figure 5.1 – Python script for keyword similarity between verses

Analysing the results show that considering only keywords in similarity analysis still leaves many similar verses undetected because they share 'concepts' and not 'keywords'. This makes yet another strong case for this research considering added layer of conceptual and semantic annotation. And here is a sample output of the above program.

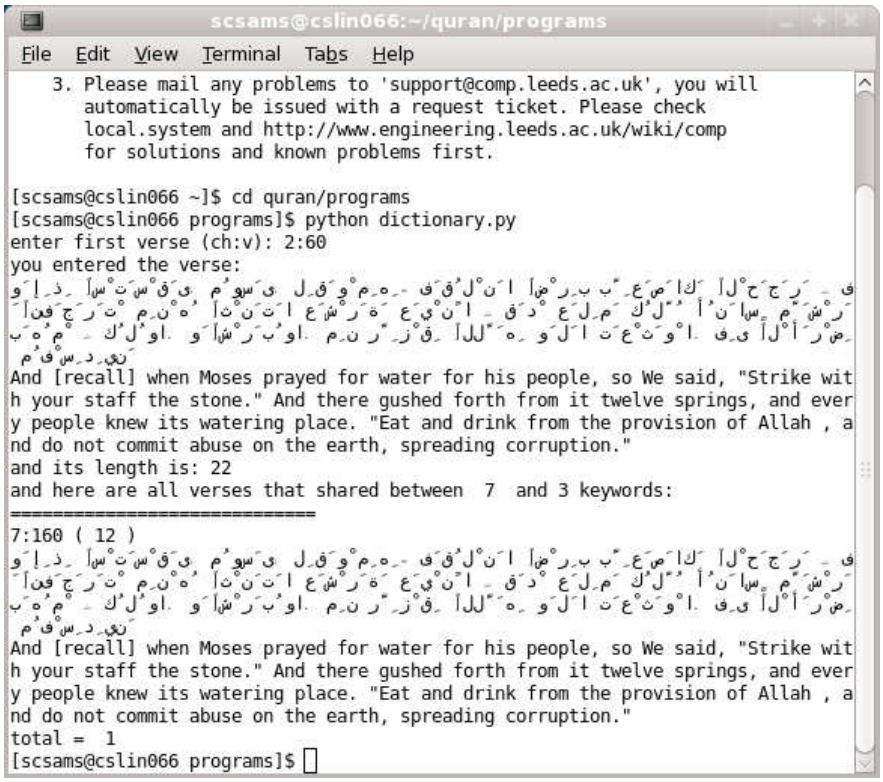


Fig 5.2 – Sample output of the similarity application

This function was modified slightly to produce a long listing of similar verses for the entire Qur'an. Following is a small snapshot, the first verse is the source, and subsequent verses are the similar verses. Note that many verses have no similarity according to our criteria.

```
67:28 ,
16:104,42:21,6:40,6:47,11:26,58:4,7:73,2:10,9:79,29:23,9:90,5:94,5:73,28:71,46:10,46:3
1,22:25,14:22,24:63,2:174,6:46,3:32,35:40,3:77,46:4,10:59,3:177,41:52,9:61
67:29 , 28:85,34:24
6:79 , 9:36,6:14
6:78 ,
28:22,36:20,6:77,6:76,12:37,71:2,20:90,61:5,11:63,11:61,5:20,11:28,2:54,27:40,27:46,18
:98,18:95,7:73,11:84,11:88,20:52,40:29,19:47,12:28,12:23,26:188,21:4,2:258,20:86,6:19,
11:54,11:50,11:78,7:85,7:65,7:67,7:61,11:92,10:15,59:16,26:62,6:80,10:71,12:50,43:51,1
2:98
6:71 , 2:120,40:66,3:73
6:70 ,
6:73 ,
6:72 ,
6:75 , 7:185
6:74 , 43:26,21:54,7:60,36:47
37:16 ,
37:17 ,
37:14 ,
37:15 ,
37:12 ,
6:77 ,
28:22,6:78,6:76,12:37,11:63,11:61,11:28,27:40,18:98,18:95,11:88,20:52,19:47,12:23,26:1
88,21:4,2:258,11:92,10:15,26:62,6:80,12:50,12:98
37:10 ,
37:11 ,
6:76 ,
28:22,6:78,6:77,12:37,12:33,11:63,11:61,11:28,27:40,18:98,18:95,11:88,20:52,19:47,12:2
8,12:23,26:188,12:4,21:4,2:258,11:92,10:15,26:62,6:80,12:50,12:98
37:18 ,
37:19 ,
84:8 ,
84:9 ,
```

5.5- MAKKI AND MADANI CLASSIFICATION USING WEKA

After experimenting with basic frequency distributions, and after gaining access to a tag corpus, I wanted to experiment with machine learning algorithms. The problem I chose is learning classifiers to distinguish between Makki and Madani chapters. The main challenge here is to choose the attribute set such that their number is optimal and at the same time considered good distinguishing criteria. From the empirical statistics found in the previous section, I can suggest few attributes like most frequent keywords found only in Makki or Madani category. However, suggesting attributes based on empirical findings comes at a second level after exploring domain knowledge. Hence, I researched works of early Qur'anic scholars on this matter. Good source of information can be found in (As-Soyouti 1987) and (As-Sabt 1996) as well as in http://www.qurancomplex.com/ .These scholars did suggest a number of distinguishing features of Makki and Madani chapters, which confirm my empirical findings. Based on my experiments and these scholarly comments, I came up with the following list of attributes summarized in the table below .

	Attribute/keyword	Classify as	Total found	Corpus Search term
1	Reference to the lemma "prostration [سجد]"	Makki	92	ROOT:sjd
2	Reference to the word "Never كلا" as an aversion particle	Makki	31	POS:AVR LEM:kal~aA"
3	'O mankind'	Makki	20	يا أيها الناس
4	'O you who believe'	Madani	89	يا أيها اللذين آمنوا
5	Initial letters	Makki	30	POS:INL
6	Prophets names	Makki	581	["<ibora`hiym", "<isoma`Eiyl", "yaEoquwb", "<isoHa`q", "muwsaY", "EiysaY","daAwud","nuwH", "zakariy~aA", "yaHoyaY", "yuwnus", "ha`ruwn", "sulayoma`n", "yuwsuf", "<iloyaAs", "yasaE", "luwT", "Sa`liH", "huwd", "Adam", "ṢuEayob", "<idoriys"]
7	Reference to the story of creation	Makki	5	Find both "Adam" and "<iboliys"
8	Use of emphasis, exhortation, aversion and certainty	Makki	1478	"CERT", "SUP", "EXH", "AVR", "EMPH"
9	Average length of verses, shorter verses are Makki		Avr: 10. 32	For each chapter: Count total words and divide by total verses
10	Reference to hell, fire, paradise, day of judgment	Makki	822	"jahan~am", "LEM:jan~ap", "naAr", "saEiyr","qiya`map", "Ea*aAb","aAxirap"
11	Reference to Jihad and fighting	Madani	211	"ROOT:qtl", "ROOT:jhd"
12	Reference to marriage, divorce, women and wife	Madani	181	"ROOT:nkH","ROOT:Tlq", "LEM:zawoj", "LEM:nisaA"
13	Reference to Jews, Christians, bible, children of Israel	Madani	97	"LEM:<isora`^", "ROOT:yhwd", "LEM:t~aworaY`p", "<injiyl", "LEM:naSoraAniy"
14	Pillars of Islam: prayer, fasting, zakat and hajj	Madani	133	"Salaw`p", "zakaw`p", "LEM:SiyaAm", "LEM:Haj~"
15	Word count	N/A	77909	N/A

Table 5.14 – Attribute Set to learn classifiers for Makki and Madani Chapters

Following is a discussion justifying the selection and classification of these attributes.

The first attribute –i.e., Prostration- is an act of worship in Islam that involves placing the most honourable part of once body –i.e., the forehead- on the ground as a symbol of submission to one God. The people of Makkah used to refuse performing this humiliating act, and hence Makka chapters repeatedly urge the Meccan to submit fully to Allah through prostration. Following is a sample verse

from 25:60 "*And when it is said to them, "**Prostrate** to the Most Merciful," they say, "And what is the Most Merciful? Should we **prostrate** to that which you order us?" And it increases them in aversion*." Verses were searched for root word ROOT:sjd which captures various inflectional forms of this verb.

The second attribute is based on our empirical observation –supported by scholarly observation as well- where the aversion particle "kal~a" meaning 'never' or 'no' is used only in Makki chapters, like the following verse 70:39: "***Never**! Indeed, We have created them from that which they know*." This particle is used in the dialogue with people of Mecca arguing over their denial of submission to God and their denial of the day of Judgement.

The third attributes are not exclusively dedicated to Makka or Median, but it is based on majority of cases where the vocative expression "O mankind" followed by some message is more often a feature in Makka chapters –but also mentioned in few Medina chapters like in 2:21- when most of the people were not yet believers. Here is an example from 10:57 " ***O mankind**, there has to come to you instruction from your Lord and healing for what is in the breasts and guidance and mercy for the believers.*"

However, later in Medina, the Muslim population grew and Qur'an started to address them starting with the expression in the fourth attribute: "O you who believe!" which appears only in Medina chapters. Here is an example of such a verse 61:10: "***O you who have believed**, shall I guide you to a transaction that will save you from a painful punishment?"*

There are 29 chapters initialized with letters, which are tagged as INL in the QAC. All except chapters 2 and 3 are Makki. We consider this as our fifth attribute. Here is an example from 3:1-2 " ***Alif, Lam, Meem**. Allah - there is no deity except Him, the Ever-Living, the Sustainer of existence.*"

The sixth attribute is based on stories of previous prophets and messengers mentioned in the Qur'an like for example story of Moses, Abraham, Noah, etc. According to scholars, these stories in the Qur'an server two purposes: first, a warning to the people of Makka that if they reject Prophet Muhammad some punishment will befall on them as it happened to previous people who rejected their messenger like Noah, Moses, etc. Second, these stories add motivation and steadfastness to Prophet Muhammad who often gets frustrated when people of Makkah continue to deny his message. In both cases, the subject is highly related to Makkah and hence this could be a good target of attribute. Chapter 2 –which is Madani- is an exception where stories of Moses and Abraham were mentioned, but in all other instances these stories always occur in Makki chapters. See for example 41:13 warning people of Makkah: "But if they turn away, then say, "I have warned you of a thunderbolt like the thunderbolt [that struck] 'Aad and Thamud." In my search, I included a number of prophets names which are mentioned in the Qur'an except Prophet Muhammad.

The seventh attribute is about a special story, which is the story of creation when Adam was created and Allah ordered angels to prostrate to Adam, all submitted except Iblis (Satan) refused. This story was mentioned only five times and our search produced a Boolean YES, NO when both names (Adam and Iblis) are mentioned in the chapter. These instances were all Makki again except chapter 2 which is Madani.

The eighth attribute is based on the rhetorical style of Makki chapters where language of certainty, surprise, exhortation, aversion, emphasis is used in the course of arguments with the people of Makkah. The QAC tags certain particles with 'emphasis', 'certainty', 'surprise', 'aversion', 'emphasis', and I exploited this feature in counting for this attribute. Hence, my results are drawn totally on usage of these particles, but later when semantic roles will be annotated the results would be more accurate. Following table gives example verses of these particles.

QAC tag	particle	Example verse
AVR (aversion)	كلا	70:39 "***Never**! Indeed, We have created them from that which they know*."
CERT (certainty)	قد	6:97 And it is He who placed for you the stars that you may be guided by them through the darknesses of the land and sea. **Certainly**, We have detailed the signs for a people who know.
SUP (Surprise)	إذاً	17:73 And indeed, they were about to tempt you away from that which We revealed to you in order to [make] you invent about Us something else; **and then** they would have taken you as a friend.
EXH (exhortation)	لولا	10:20 And they say, "**Why** is a sign not sent down to him from his Lord?" So say, "The unseen is only for Allah [to administer], so wait; indeed, I am with you among those who wait."
EMPH (emphasis)	لام التوكيد	11:10 But if We give him a taste of favor after hardship has touched him, he will **surely** say, "Bad times have left me." Indeed, he is exultant and boastful

Table 5.15 – Example verses of certainly and emphasis particles in the Qur'an

The ninth attribute is straightforward and deals with average length of a verse in terms of words. Makkah verses are shorter than Medina verses which tends to be long when discussing legislations and rulings.

The tenth attribute is counting the reference to some unseen facts and events from the world hereafter which the people of Makkah used to deny and hence Makki chapters repeatedly emphasized on them. Search was made on few concepts like Hellfire –and some alternative names of hellfire in the Qur'an, punishment, paradise and the word "aAxirap" meaning hereafter.

Military conflict with the people of Makkah happened in Islamic history only after the migration of Prophet Muhammad to Medinah, and hence our eleventh attribute on "jihad" in its 'fighting' sense appears only in Medina chapters. Searching for this attribute was done through the root word of 'qtl' meaning 'to fight' and 'jhd' meaning to 'struggle'. This is a near approximation as some reference to struggle by earlier prophets might wrongly be included, nevertheless this gives a close estimation.

The twelfth attribute is about the concept of family legislations like marriage, divorce, pregnancy, breast feeding, etc. These details are only mentioned in Medina chapters. Our search included root word for 'marriage' and 'divorce', also we included the lemma 'wife' and 'women'. The latter two lemmas would bring in some false positives where 'zawoj' is used to mean a 'pair' of anything like the word 'mates' in 51:49 which is a Makki chapter: "And of all things We created two **mates**; perhaps you will remember."

Attribute number 13 searches for reference to other divine religions mainly Jews and Christians. It is only when Prophet Muhammad migrated to Medina he encountered with Jew tribes who used to live in Medina, and later some Christian delegates came to Medina to debate on the nature of Jesus. However, reference to Jesus is made in a number of Makki chapters, and this is why in my search I included only terms: 'children of Israel', 'Torah', 'Gospel', 'Jew' and 'Christian'.

Apart from the first pillar of Islam all the other four (i.e., prayer, fasting, zakat and Hajj) were all obligated in Medina period. Hence, I searched for lemma on these four pillars. Again there could be false positives like the following Makki verse 19:55 referring to Prophet Ishmael: "And he used to enjoin on his people prayer and zakah and was to his Lord pleasing."

After designing this attribute set, a spreadsheet is populated with these data and converted into weka 'arff' file. Following is a snapshot of the 'arff' file.

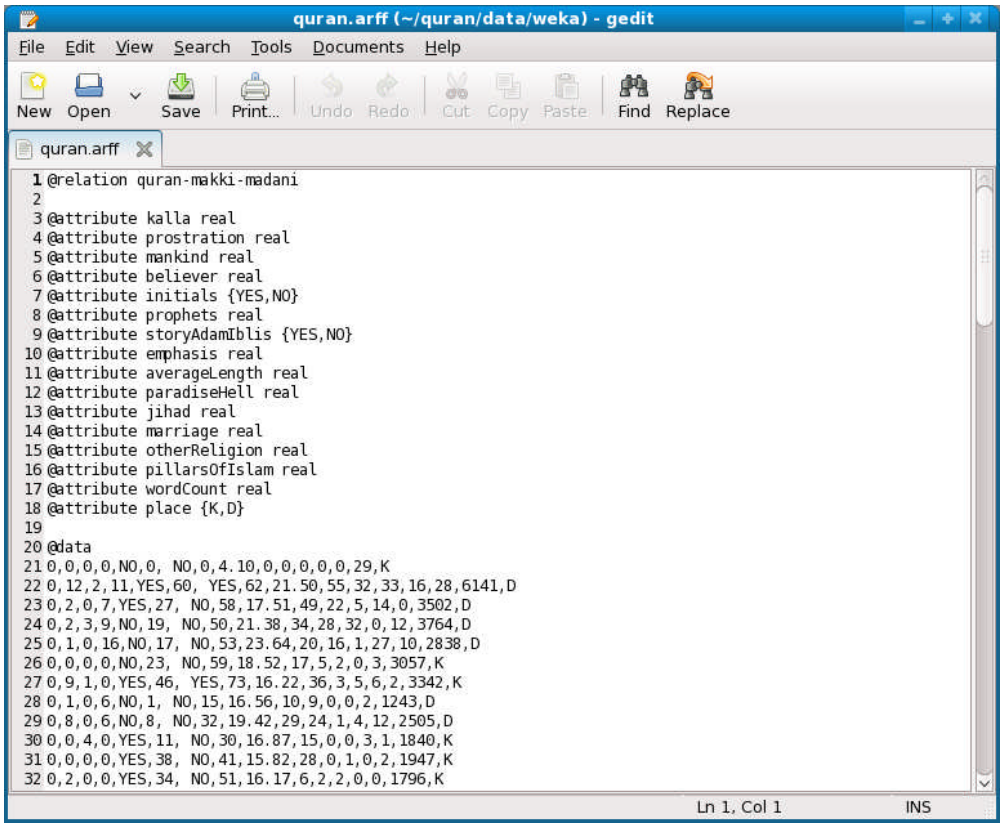


Fig 5.3 – 'ARFF' file, each line of data represents one chapter starting from chapter 1 to 114

Weka created a model for this data as shown in the figure below.

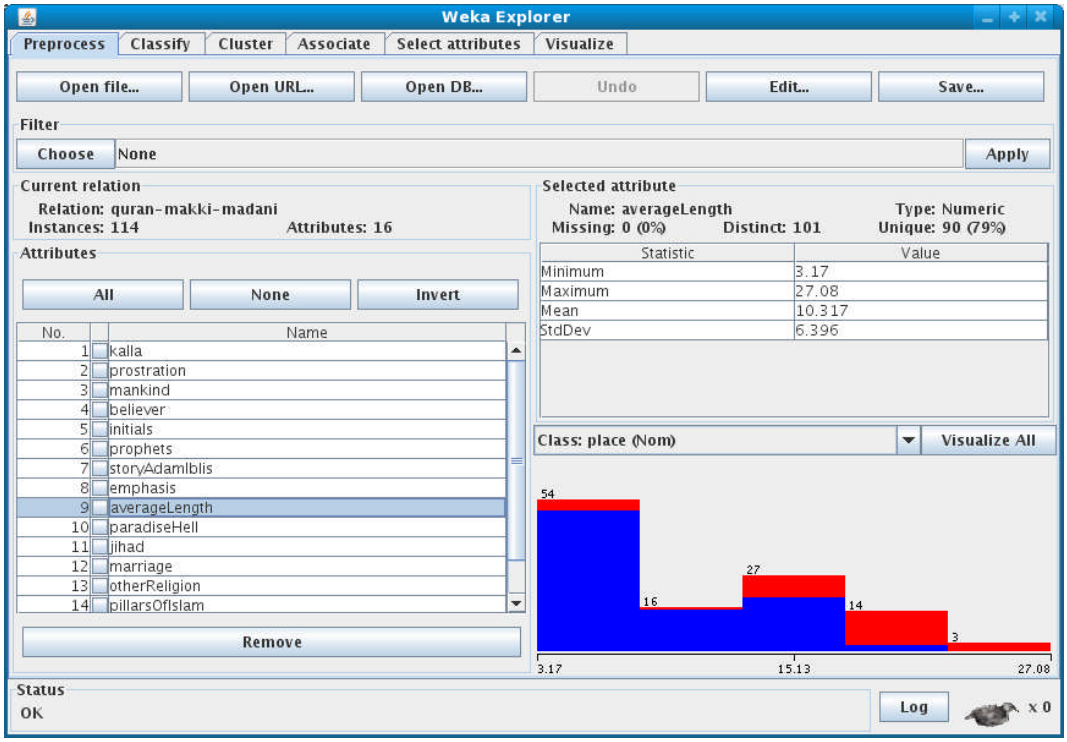


Fig 5.4 – Weka Model for Makki [blue] and Madani [red] chapters

Weka enabled various convenient visualization of these data. Figure 5.4 (next page) gives a plot of these 16 attributes.

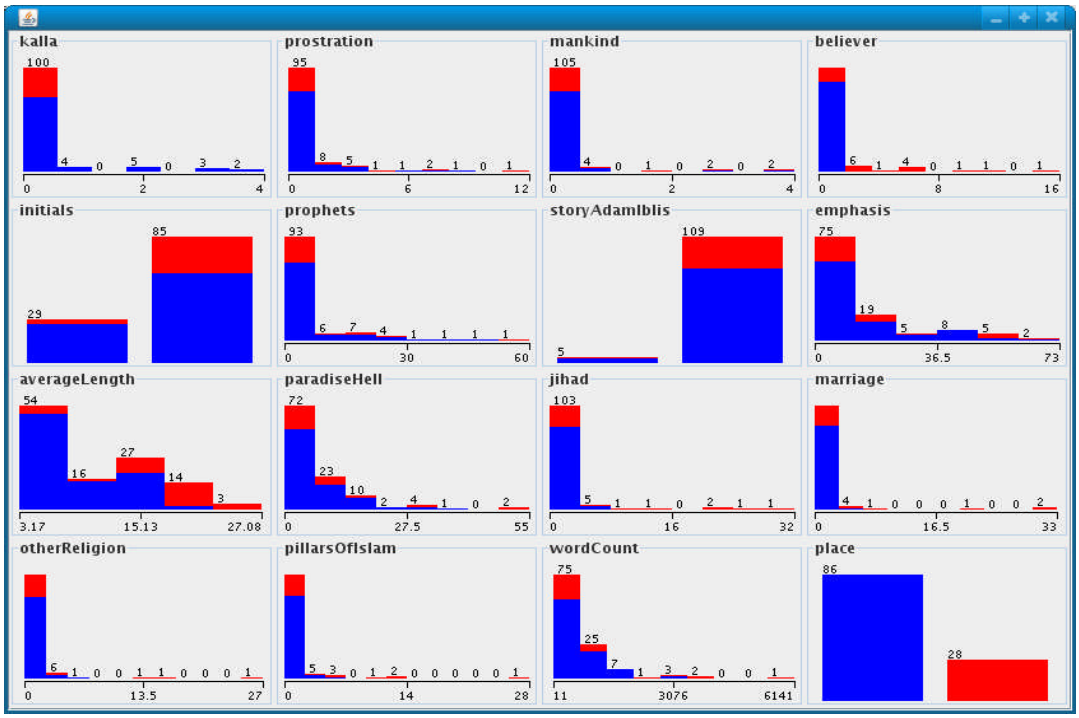


Fig 5.5 – Visual Plot of all 16 attributes, BLUE is Makki [K], and RED is Madani [D]

Next, C4.5 classifier is used to create the following decision tree.

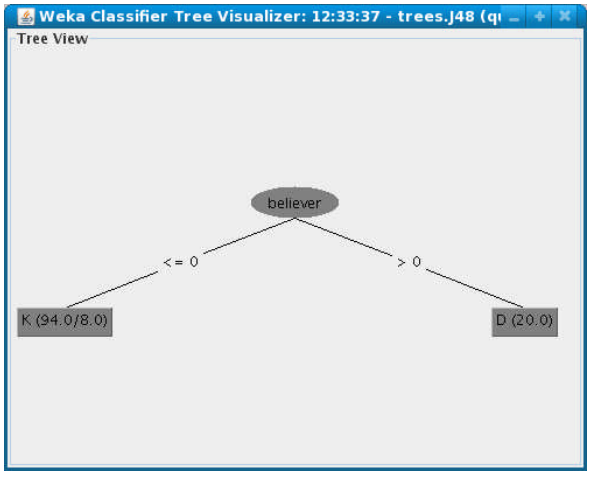


Figure 5.6 – Decision tree using C4.5 classifier

Following is the summary result of the classification:

Correctly Classified Instances	104	91.2281 %
Incorrectly Classified Instances	10	8.7719 %
Kappa statistic	0.7449	
Mean absolute error	0.1297	
Root mean squared error	0.2866	
Relative absolute error	34.758 %	
Root relative squared error	66.544 %	
Total Number of Instances	114	

I experimenting with other classifiers, and only REP Tree produced more fine grained decision tree as follows:

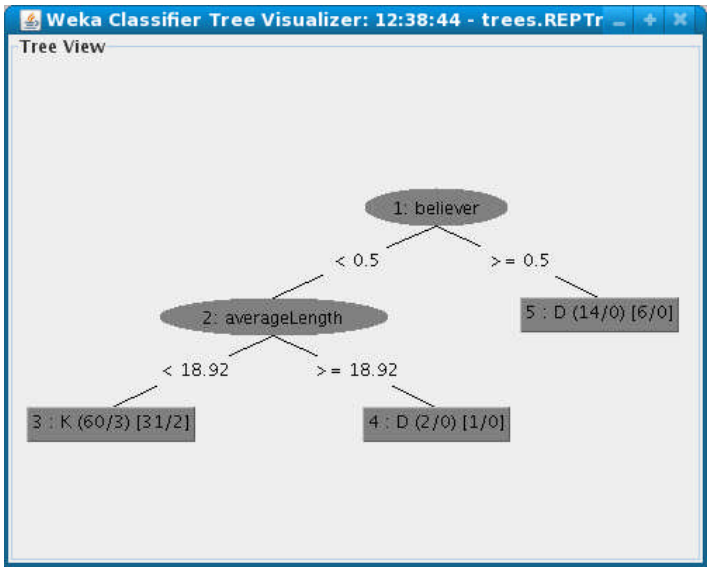


Fig 5.7 – REPTree classifier Decision Tree

And is the summary analysis of this classifier.

Correctly Classified Instances	106	92.9825 %
Incorrectly Classified Instances	8	7.0175 %
Kappa statistic	0.7904	
Mean absolute error	0.1222	
Root mean squared error	0.2571	
Relative absolute error	32.746 %	
Root relative squared error	59.6798 %	
Total Number of Instances	114	

It is noted that most Makki chapters are short and often show no occurrence of the attribute resulting is many 'zero' values. To overcome this problem, I normalized the data by adding one and dividing the absolute count by total number of words for that chapter. With that I produced another normalized model. Here are the attribute plots after normalization.

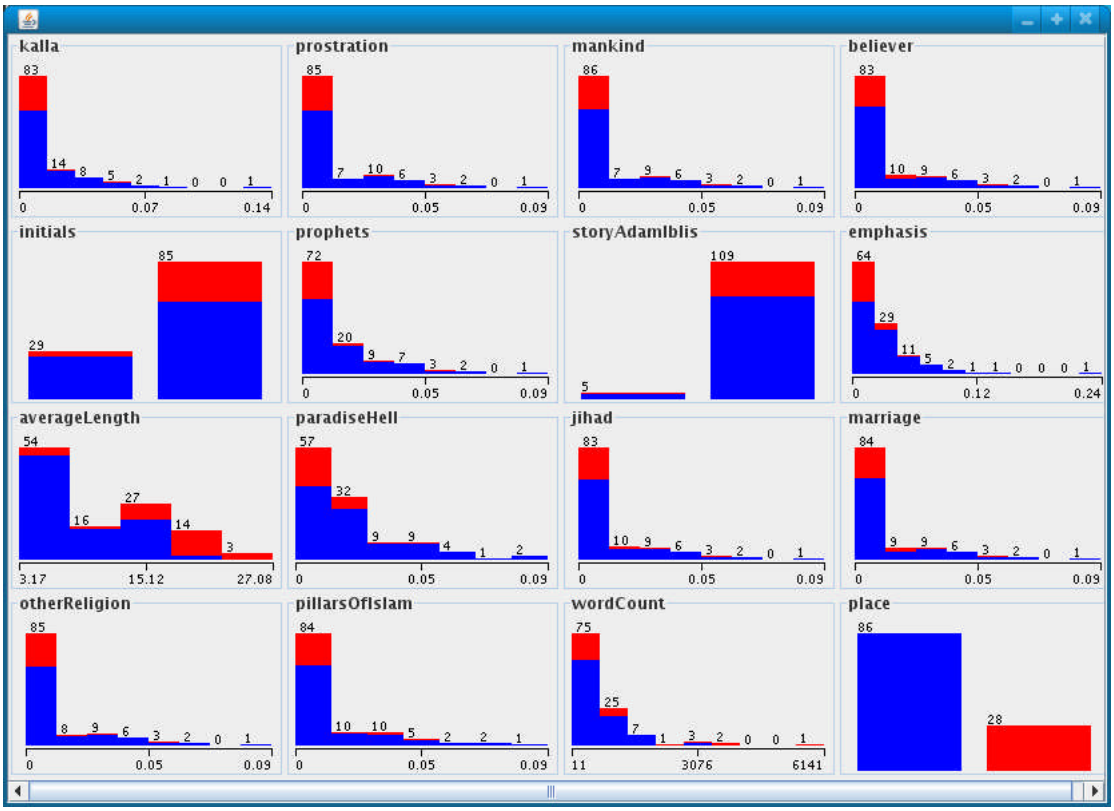


Fig 5.8 – Attribute plots after normalization, RED=Madani[M], NO, and BLUE=Makki[K], YES

I re-run the C4.5 classifier and decision tree now looks better although the accuracy was less.

```
Correctly Classified Instances          95               83.3333 %
Incorrectly Classified Instances        19               16.6667 %
Kappa statistic                          0.4956
Mean absolute error                      0.1853
Root mean squared error                  0.3808
Relative absolute error                 49.6296 %
Root relative squared error             88.4005 %
Total Number of Instances              114

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.942     0.5        0.853     0.942     0.895    K
  0.5       0.058      0.737     0.5       0.596    D

=== Confusion Matrix ===

  a  b   <-- classified as
 81  5 |  a = K
 14 14 |  b = D
```

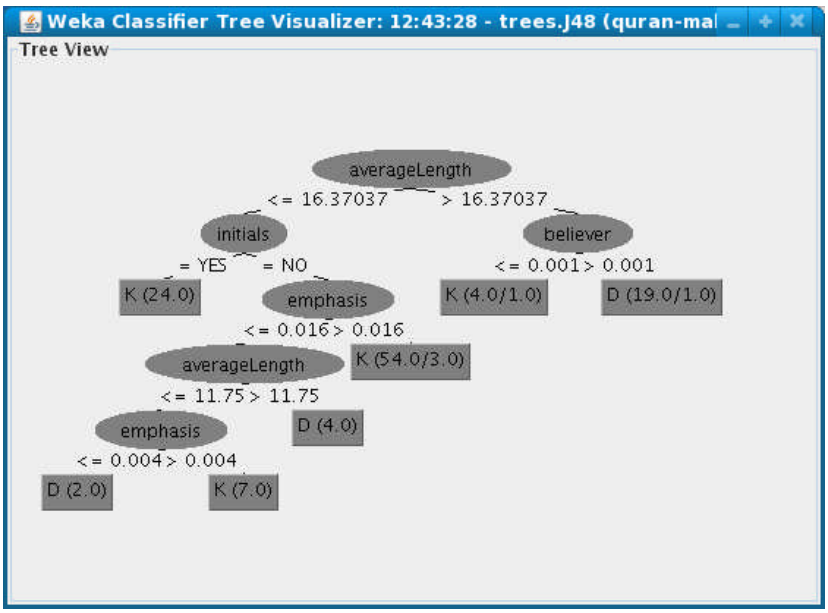


Fig 5.9 – C4.5 Decision tree after normalization

From this experimentation we can note the following observations and areas of improvement, which will be carried out in my next year research:

1. Richer annotation of the Qur'an is likely to produce better results. As noted we only leveraged on keywords, morphological and POS features in classifying between Makki and Madani. Any ontological or semantic annotation would have produced much better results.
2. Machine learning algorithms produced interesting empirical findings that would interest Qur'anic scholars, for example from figure 5.7, we came to know that verses in Makki surah's are 18 words in average.
3. The convenient visualization in Weka enables quick validation of certain observations made by early scholars, for example figure 5.5 shows that there are Madani chapters (red) that

contain the construct “O Mankind”, and this refute claims by some scholars that this construct appears only in Makki chapters.

4. This experiment showed which attributes are likely to help learning classifiers, for example, only 4 attributes among the 15 were chosen by our classifiers: believer, emphasis, averageLength and initials.
5. This experiment can be further improved by considering verses instead of chapters. This would result in defining 6240 lines of data. The obvious problem will be the abundant ‘zero’ values for certain attributes.
6. Although this experiment showed success in machine learning for classifying Makki and Madani verses, other text mining problems could be modelled similarly, for example, finding verses that are similar, and finding associations and patterns. This experiment acted as a feasibility study which will further be verified in the pilot project.

CHAPTER 6 – CONCLUSION

This report gave detailed account on my research work so far and my plan in future. I believe my research topic of text mining the Qur'an is novel and would benefit a wide range of audience. There is a huge potential in researching this area, and my research will contribute towards the growing field of Arabic NLP and Arabic corpus linguistics. This project's key contribution is in building the appropriate resources for text mining the Qur'an. This project is motivated by the increasing interest on our target text (i.e., the Qur'an), and the lack of resources and researches in this field, added to it the availability of many scholarly commentaries on the Qur'an makes it easier to develop annotation guidelines and resolve conflicts.

In what follows I would like to rate myself against the eight evaluation criteria in the assessment form.

Research Focus: Focusing my project towards the Qur'an is sensible for two reasons: first because of a wide interest in this text, and second, because of its comparatively small size annotation works would be feasible within a PhD timeline. The acceptance of a work-in-progress paper in the Corpus Linguistics conference-2009 with high rating is an initial sign of interest in this research. This project is first of its kind in the extent computational methods are used to investigate the Qur'an. This work should motivate future research on many other related areas. Based on these, I believe I should score 1 on this criterion.

Research Methodology: I have divided the project into a pilot annotation considering chapters 2 and 7 from which I will consider the most interesting features in annotating the entire Qur'an. I have adopted the CRISP-DM methodology of data mining in my research steps, and developed a realistic plan. This industry standard methodology is likely to keep the project in track. I have identified areas where text mining differs from data mining and made appropriate amendments to CRISP-DM to cater for such changes. Given these, I believe my research methodology is well-defines, appropriate and justified and hence should score 1 as well.

Originality: Apart from few papers discussed in chapter 2 which employs computational empirical methods in distinguishing between Makki and Madani chapters, I could not locate academic researches on multi-level annotation for text mining the Qur'an. I believe thus this is the first attempt of its kind considering Qur'an as the domain, and hence would deserve scoring 1 against this criterion.

Progress: I have demonstrated my competency in applying computational and machine learning tools to my domain (Chapter 5). I have demonstrated how a POS tagged corpus improved the results over the raw text, and it is the intention of this project to improve results further through added layers of annotation. I have managed to successfully publish work-in-progress paper on my research and present it in front of experts in this field.

Myself as a researcher, I feel very comfortable with the domain (i.e., the Qur'an) being an expert in classical Arabic language and the content of the Qur'an as well as scholarly comments. Moreover, I have memorized the Qur'an by heart from my early childhood, and hence can recall easily the verses under investigation. Also, my six years experience in industry as a software quality assurance officer will also speak in favour of producing methodology based quality output. during the past six months, my focus changed where initially I considered developing FrameNet type of role labelling, then considered studying the 'particles and prepositions', and then considered 'QA system', before arriving to the present scope detailed in this report. Overall I think I should score between '1' and '2' here.

Literature Review: I have investigated a wide range of literature both considering the Arabic sources on the Qur'an, Quranic science and Tafsir, as well as current trends and achievements in the field of

text mining. I became acquainted with domain specific works on biological and medical text mining, as well as computational stylistics of literature, and I have discussed how advancement in these fields can benefit my research on Qur'anic annotation and text mining. In total I have included more than 80 citations. I should score '1' against this criterion.

Report – Work Done: Chapters 4 and 5 are dedicated on reporting my work done so far. I have kept a copy of the CL2009 paper and the submitted abstract of LREC2010 in the appendix. I obtained a copy of the Arabic Quranic Corpus lately when it was recently released, and had it been available earlier, more computational work would have been possible to demonstrate. Here, I believe I should score '1'.

Report – Research Plan: Chapter 3 gives a detailed description of my research plan. I have included a Gantt chart and considered various opportunities of publishing my work. I described the project according to the phases of the CRISP-DM methodology. To be realistic I considered a pilot annotation of the Qur'anic considering only chapters 2 and 7. I believe my plan is credible and realistic, and hence would deserve score '1'.

Training Requirements: I have included in Chapter 4 a list of courses I joined. My annotation work on Arabic discourse Treebank (under Katja Markert and Amal As-Saif) contributes towards my training needs. I could see still opportunities of improvement especially in mathematical and statistical background for machine learning, as well as some GUI programming under Java in case I need to build annotation tools. I think I should score '2' under this criterion.

REFERENCES

Ad-Dinawry, A. M. (died in 855) (1978). Tafsir Gareeb al-Quran. [تفسير غريب القرآن] Dar al-Kutub al-Elmiyyah. [http://www.qurancomplex.com/TafGhreeb.asp]

Agrawal, R., Imielinski, T., and Swami, A. (1993). "Mining Association Rules between Sets of Items in Large Databases". In Proc. of ACM SIGMOD.

Al-Baghawi, A. M. (died in 1089) (1997). Ma'alim at-Tanzeel [معالم التنزيل]. Dar Taibah lin-Nashr wat-Tawze'a, 4th Edition. Also available online at: [http://www.qurancomplex.com/Quran/tafseer/Tafseer.asp?t=BAGHAWI]

Al-Kharrat, A M. (2007) Eraab Mushkil al-Quran (إعراب مشكل القرآن) Syntactic Analysis of the Quran. Also available online at [http://www.qurancomplex.com/earab.asp]

Alm, C. O., Roth, D., and Sproat, R. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (Vancouver, British Columbia, Canada, October 06 - 08, 2005). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 579-586

Al-Qahtani, D (2005) Semantic Valence of Arabic Verbs. Librairie du Liban Publishers

Ananiadou, S., and McNaught, J. (eds) (2006). Text Mining for Biology and Biomedicine. Artech House.

As-Sabt, K. A., (1996) Qawaed at-Tafseer [قواعد التفسير]. PhD thesis, Islamic University of Medinah.

As-Soyouti, J. (died in 1490) (1987) Al-Etqan fi Uloom al-Quran. [الإتقان في علوم القرآن] Dar Ibn Kathir, Damascus.

At-Tabari, M. J. (died in 889) (2000). Jeme' al-Bayan fi Ta'weel al-Quran [جامع البيان في تأويل القرآن]. Moassat al-Risalah. also available online at [http://www.qurancomplex.com/Quran/tafseer/Tafseer.asp?t=TABARY]

Banfield, A., (1982) Unspeakable Sentences: Narration and Representation in the Language of fiction. Routledge and Kegan Paul.

Blake, C., and Pratt, W. (2001)."Better rules, fewer features: a semantic approach to selecting features from text". In Proc. of the 2001 IEEE Int conference on Data mining. San Joes, CA, IEEE Computer Society press, New York:59-66.

Bloehdorn, S., and Hotho, A. (2006). "Boosting for text classification with semantic features". Lecture Notes in Computer Science. Springer.

Bodenreider, O. (2006). "Lexical, Terminological, and Ontological Resources for Biological Text Mining". In Text Mining for Biology and Biomedicine.

Chew,P.A., Verzi,S., Bauer,T. and McClain,J. (2006)."Evaluation of the bible as a resource for cross-language information retrieval". In Proceedings of the Workshop on Multilingual Language.

Church, K. W., and Mercer, R. (1993). "Introduction to the Special Issue on Computational Linguistics Using Large Corpora." *Computational Linguistics,* 19(1): 1–24.

Cohen, W.W., and Singer, Y. (1999). "Context-sensitive learning methods for text categorization". ACM trans. Inform. Systems. 13(1): 100-111.

Cook, J. (ed) (2002). Bible and Computer. *The Stellenbosch AIBI-6 Conference*. Brill.

Cook, W. (1979) Case Grammar: Development of the Matrix Model. Washington, D.C. : Georgetown University.

Dagan, I., Karov, Y. and Roth, D. (1997). "Mistake-driven learning in text categorization". In Proc. of EMNLP-97, pp 55-63.

Darwish, M. (1999) Earab al-Quran wa Bayanihi [إعراب القرآن وبيانه], Dar Ibn Katheer, Damascus.

Dhillon, I., Mallela, S., and Kumar, R. (2002). "Enhanced Word Clustering for Hierarchical Text Classification". In Proc. Of KDD-02.

Dror J., D. Shaharabani, R.Talmon and S. Wintner. (2004) Morphological Analysis of the Qur'an. *Literary and Linguistic Computing*, 19(4):431-452

Dukes, K and N. Habash (2010). *The Crescent Corpus: Morphological Annotation of the Holy Quran*. Submitted to the seventh international conference on Language Resources and Evaluation (LREC-2010). Valletta, Malta.

Dukes,K., Atwell, E. and Sharaf, A. (2010) Syntactic Annotation Guidelines for the Quranic Arabic Treebank. Submitted to the seventh international conference on Language Resources and Evaluation (LREC-2010).

Dumais, S. T., Platt, J., Heckerman, D. and Sahami, M. (1998). "Inductive learning algorithms and representations for text categorization". In Proc. of CIKM-98, 7th International Conference on Information and Knowledge management (Bethesda, MD), 148-155.

Ekman, P. (1982) *Emotion in the Human Face*. Cambridge University Press, Second ed.

Elkateb, S., Black. W., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C. (2006) Building a WordNet for Arabic, in Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.

Esuli, A., and Sebastiani, F. (2006). "SentiWordNet: A publicly available lexical resource for opinion mining". LREC 2006.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popsecu, A., Shaked, T., Soderland, S., Weld, D., Yates, A. (2005). " Unsupervised named entity extraction from the web: An experimental study." Artificial Intelligence 165(1): 91-134.

Fillmore, C.(1968) The case for case. In Bach, E. W. and Harms, R.T. (Eds), Universals in Linguistic Theory.

Fiteih, M. (1983) Prepositions and Prepositional Verbs in Classical Arabic. PhD thesis, Univesity of Leeds.

Fledman, R., and Sanger, J. (2007). "The text mining handbook". Cambridge press.

Godbole, N., Srinivasaiah, M. and Skiena, S. (2007) "Large-scale sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Hayes, P. (1992). "Intelligent high-volume processing using shallow domain-specific techniques". In Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval.

Holmes, D. (1992)."A stylometric analysis of Mormon scripture and related texts". Journal of Royal Statistical Society, Series A, 155(1):91-120.

Hotho, A., Staab, S., and Stumme, G. (2003). “Text Clustering Based on Background Knowledge”. Institute of Applied Informatics and Formal Descriptive Methods, University of Karlsruhe, Germany: 1-35.

Hu, M., and Liu, B. (2004). Mining and Summarizing Customer Reviews. In Procs. of KDD, Seattle, WA.

IbnKatheer, I. O., (died in 1353) (1999).Tafsir Al-Quran Al-Adhym [تفسير القرآن العظيم]. Dar Taibah lin-Nashr wat-Tawze'a. Also available online at: [http://www.qurancomplex.com/Quran/tafseer/Tafseer.asp?t=KATHEER]

Joachims. T. (1998). "Text categorization with support vector machines: learning with many relevant features." In Proc. of ECML-98, 10th European conference on Machine Learing. (Chemnitz, Germany, 1998), 137-142.

Kim, J., and Tsujii, J. (2006). “Corpora and their Annotation”. In Text Mining for Biology and Biomedicine.

Kim, S.M., and Hovy, E. (2007). "Crystal: Analyzing predictive opinions on the web," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* , 2007.

Lam, W., and Ho, C. Y. (1998). "Using a generalized instance set for automatic text categorization". In Proc. of SIGIR-98.

Laver, M., Benoit, K. and Garry, J. (2003) "Extracting policy positions from political texts using words as data," *American Political Science Review*, vol. 97, pp. 311-331

Lin, D. (1998). "Dependency-based evaluation of MINIPAR." In Proc. of ICLRE'98 worshop on Evaluation of Parsing Systems.

Liu, H., Lieberman, H., and Selker, T.(2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international Conference on intelligent User interfaces* (Miami, Florida, USA, January 12 - 15, 2003). IUI '03. ACM, New York, NY, 125-132

Mahlberg, M. (2007). "Clusters, key clusters and local textual functions in Dickens". *Corpora*, 2(1): 1-31.

Martin, L. W. and Vanberg, G. (2008) "A robust transformation procedure for interpreting political text," *Political Analysis*, vol. 16, pp. 93-100, 2008.

McNaught, J. And Black, W.J. (2006). “Information Extraction”. In Text Mining for Biology and Biomedicine.

Mir, M (1989) Verbal Idioms of the Quran. Michigan Series on the Middle East, No. 1.Center for Near Eastern and North African Studies, University of Michigan, Ann Arbor.

Mullen, T. and Malouf, R. (2006) "A preliminary investigation into sentiment analysis of informal political discourse," in *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 159-162.

Müller, C., and Strube, M. (2006): Multi-Level Annotation of Linguistic Data with MMAX2. In: Sabine Braun, Kurt Kohn, Joybrato Mukherjee (Eds.): Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt: Peter Lang, pp. 197-214. (English Corpus Linguistics, Vol.3)

Pang, B., and Lee, L. (2008)."Opinion Mining and Sentiment Analysis". Foundations and Trends in Information Retrieval. (2):1-135

Park, J. C., and Kim, J. (2006). "Named Entity Recognition" In Text Mining for Biology and Biomedicine.

Piasecki, M., Szpakowicz,S., Broda, B. (2009). A Wordnet from the Ground Up. Wroclaw.

Plaisant, C., Rose, J., Yu, B., Auvil, L., Kirschenbaum, M. G., Smith, M. N., Clement, T., and Lord, G. (2006). Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces. *In Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (Chapel Hill, NC, USA, June 11 - 15, 2006). JCDL '06. ACM, New York, NY, 141-150.

Popescu, A. M. and Etzioni, O. (2005) "Extracting product features and opinions from reviews". In Procs. of HLT-EMNLP.

Radev, D. R., Hovy, E., and McKeown, K. (2002) "Introduction on the special issue of summarization". Computational Linguistics, 28:399-408.

Resnik, P., M. B., Olsen, and Diab, M. (1999). "The Bible as a Parallel Corpus: Annotating the Book of 2000 Tongues." *Computers and the Humanities*, 33: 129-153.

Ruppenhofer,J., Ellsworth, M., Petruck,M. R. and JohnsonC. R. (2005). FrameNet: Theory and Practice.

Salih, Bahjat Al-wahhab (2007) Al-eraab al-Mufassal Li-Kitab Allah al-Murattal[الإعراب المفصل لكتاب الله المنزل] Dar Al-Fikr, Beirut

Sebastiani, F. (2002). "Machine Learning in Automated Text Categorization." ACM Computing Surveys 34(1):1-47.

Shamsan, A. (1986). Al-Fi'l fil Quran al-Kareem. (In Arabic). Kind Saud University.

Sharaf, A. and Atwell, E., (2009). A Corpus-based computational model for knowledge representation of the Quran. In the fifth corpus linguistics conference, Liverpool.

Shearer, C., (2000). "The CRISP-DM Model: The New Blueprint for Data Mining", Journal of Data Warehousing. 5(4):13-22.

Shenassa, M.E., Khalvandi, M.J. (2008). "Evaluation of different English translations of Holy Koran in scope of verb process type". Third International Conference on Information and Communication Technologies: From Theory to Applications.

Snyder, B. and Barzilay, R. (2007) "Multiple aspect ranking using the Good Grief algorithm," in *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, pp. 300-307.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000). "Automatic Text Categorization in Terms of Genre and Author." Computational Linguistics 26(4):471-495.

Starcke, B., (2006). "The phraseology of Jane Austen's Persuasion: Phraseological units as carriers of meaning". *ICAME Journal*, 30, 87-104.

Stubbs, M. (2005). "Conrad in the computer: examples of quantitative stylistic methods". *Language and Literature*, 14 (1): 5 – 24.

Subasic, P. and Huettner, A. (2001) "Affect analysis of text using fuzzy semantic typing," *IEEE Transactions on Fuzzy Systems*, vol. 9, pp. 483-496.

Swanson, D. (1990). "Medical Literature as a Potential Source of New Knowledge," Bulletin of the Medical Library Association, 78(1):29-37.

Thomas, M., Pang, B. and Lee, L. (2006). "Get out the vote: Determining support or opposition from congressional floor-debate transcripts," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 327-335.

Thabet, N. (2005). Understanding the thematic structure of the Qur'an: an exploratory multivariate approach. Proceedings of the ACL Student Research Workshop, Association for Computational Linguistics, 7-12.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In proc. of ACL.

Weiss, S. M., Apte, C., Damerau, F.J., Johnson, D. E., Oles, F. J., Goetz, T., and Hampp, T. (1999). Maximing text-mining performance. IEEE Intelligent Systems. 14(4):63-69.

Wiebe, J., Wilson, T. and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation (formerly Computers and the Humanities) 1(2)

Wiebe ,J., Wilson , T., Bruce, R., Bell, M. and Martin, M.(2004) "Learning Subjective Language". Computational Linguistics, v.30 n.3, p.277-308.

Witten, I., and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). *Morgan Kaufmann*.

Yang, Y. (1999). "An evaluation of statistical approach to text categorization." Information Retrieval Journal, 1(1): 69-90.

Yang, Y., and Liu, X. (1999). "A re-examination of text categorization methods." In Proc. of SIGIR-99, 22nd ACM International Confernece on Research and Development in Information Retrieval. 42-49.

Yi, J., Nasukawa, T., Bunescu, R., Niblack, W. (2003). Sentiment Analyzer: Extracting Sentiment about a Given Topic using Natural Language Processing Techniques. In Procs of ICDM

APPENDICES

APPENDIX 1 – LREC 2010 ABSTRACT

TOWARDS AN EFFICIENT QA SYSTEM ON THE QUR'AN THROUGH MULTI-LAYERED ANNOTATION

Abdul-Baquee M. Sharaf Eric S. Atwell

University of Leeds, UK

{scsams,scs6ea@leeds.ac.uk}

The Qur'an is believed to be the words of God and supposed to be a source of valuable information not only for a growing 1.5 billion Muslims worldwide, but also for the whole mankind. The Qur'an consists of 114 chapters with approximately 77 thousand words of classical Arabic.

The Qur'an covers a wide range of topics from description of God's attributes and qualities, to information about past historic events, to laws and legislations for Muslims, to information about future events. Thus, the Qur'an contains narrative information as well as commands and imperative statements. It was compiled over 23 years of the lifetime of the prophet Muhammad. This book is divided into varying size chapters each consists of varying size verses. Although, a particular chapter covers a unifying theme, its verses often cover a wide range of topics, which mostly correlates with related verses from different chapters. This feature of related concepts scattered within many chapters makes it a laborious human task when attempting to collect all these verses in search of a complete concept, hence, an interesting task for computational analysis.

The Qur'an was revealed on the Prophet Muhammad who in turn used to explain these verses and show to Muslims the details of generalities that was left unexplained in the Qur'an. For example, the Qur'an only commands to perform prayer, fasting, pilgrimage and so on, but the details of how, when and where to perform these obligatory are left for the Prophet to explain. Prophetic sayings are called Hadith. Moreover, the rulings and legislations within the Qur'an were gradually completed to its final shape. Thus, a later verse might come to abrogate the ruling set by an earlier verse.

All these issues and contexts need to be taken into account in order to properly understand the Qur'anic texts. Thus, when intending to understand a particular verse, all related verses need to be first consulted, then all sayings of the Prophet on any of these verses need to be referenced, along with comments of Qur'anic scholars, especially the students of the Prophet. Finally, since the Qur'an is a classical Arabic text, traditional Arabic grammar and lexicon can be helpful as well. Islamic traditional library contains many books on Qur'anic exegesis which give this kind of elaborate analysis for each verse of the Qur'an. These volumes are known as books of *Tafsir*. Famous traditional books include [At-Tabari 2000; IbnKatheer 1999 and Al-Baghawi 1997]. Extensive analysis of these books of Tafsir reveals a set of general rules for interpreting the Qur'an, which we will refer to as *Rules of*

Tafsir. One such rule specifies that any verse which addresses the Prophet Muhammad equally addresses the general Muslim community unless explicitly mentioned otherwise.

In this project, work is in progress to build a specialized multi-layered corpus of the Qur'an for the purpose of knowledge extraction and question answering. This annotation scheme is able to structure the information in Tafsir Ibn Katheer and link each verse with related verses, Hadith and scholarly comments. In order to enable computational tasks, in addition to these three specialized Tafsir layers, some more NLP layers are created. This includes: part-of-speech, syntactic parsing, FrameNet semantic role labeling, named entity resolution, anaphora resolution, and discourse relations. In the process of automatic knowledge extraction and question answering, this multi-layered annotated corpus is integrated with static knowledge base containing Arabic lexicon, Arabic grammar and rules of Tafsir. Figure 1 depicts these annotation layers and the static knowledge bases.

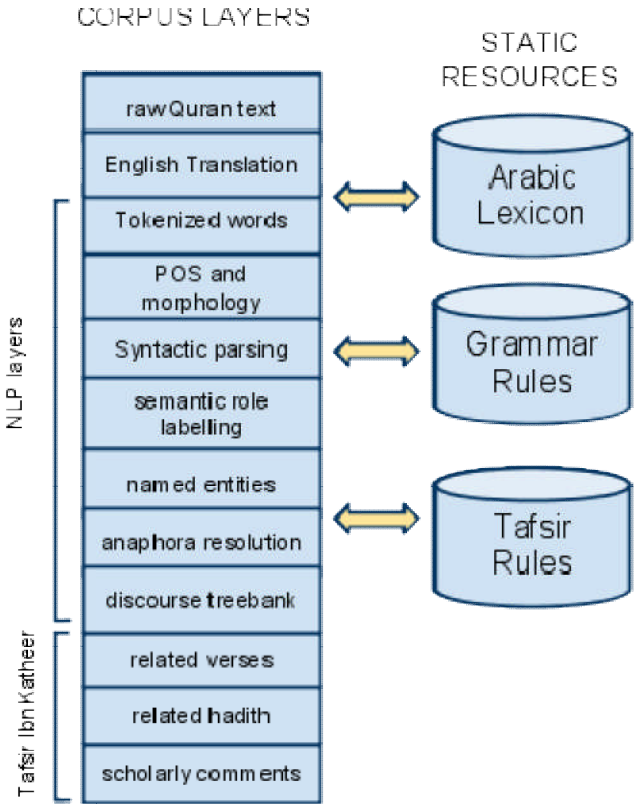


Figure 1. The Qur'an annotation layers

A pilot annotation is being carried out for chapter 2 of the Qur'an, which is the longest chapter and represents around 7% of the Qur'an covering a wide range of topics.

This research project is motivated by the fact that the Qur'an is a popular book and a source of guidance for over 1.5 billion Muslims worldwide, added to the fact that no systems exist today for knowledge extraction from the Qur'an –to the best of our knowledge- that harness the computational power, apart from online keyword (and root word) searches. See for example online Qur'an search at [http://www.islamicity.com/QuranSearch/] and [http://Quran.muslim-web.com].

Annotation guidelines for most of our NLP layers are adopted -with some customization- from publicly available resources or schemes. For morphological and POS tagged layer, there are to our knowledge two resources available: the Haifa corpus (Dror et al 2004) and the CRESENT corpus [Dukes and Habash 2010a] which we intend to use and is available at (http://www.Qur'an.uk.net). For syntactic parsing layer, we will adopt a dependency based parsing scheme incorporating traditional Arabic grammar rules described in [Dukes et al 2010b] and partially available at [http://Quran.uk.net/Treebank.aspx]. The semantic parsing layer is constructed using the annotation guideline of the FrameNet [Ruppenhofer et al 2005] which also resulted in annotating the German TIGER corpus [Burchardt et al 2009] and influenced designing a semantic annotation model for the Qur'an as described in [Sharaf and Atwell 2009]. For constructing the discourse layer we follow the guidelines in the Penn Discourse Treebank manual (Prasad et al 2008) and available at [http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf] and the explicit Modern Standard Arabic discourse annotation guideline developed at the University of Leeds [Alsaif and Markert 2009]. For anaphora resolution, we adopt the XML-based scheme described by (Tutin et al 2000) that includes coreference, set membership, substitution, sentential anaphora and indefinite relations. Rules of Tafsir have been addressed in many Arabic literatures, and in this research we adopt such rules from [As-Sabt 1996], which contains along with Tafsir rules, those rules dictated by Arabic grammar and the linguistic styles of classical Arabs. Arabic lexicon will be constructed using existing Arabic WordNet lexicon [Elkateb et al 2006] also available online at [http://www.globalwordnet.org/AWN/], and by incorporating some traditional works on Qur'anic vocabulary like [Ad-Dinawry 1978].

The annotation will be carried out through in-house developed tool, which incorporates machine readable text of the Qur'an and Tafsir of Ibn Katheer, and creates XML representation for each layer indicating the index location of the intended annotated arguments. In order to facilitate annotation, annotation scheme will be produced along with DTD description for each layer.

Having this multi-layered corpus in place, an efficient framework for information extraction and question answering can be built. The characteristics of the Qur'an portraying a network of verses sharing similar concept can be leveraged, and if only one verse can be identified from user question or input, then this verse can be considered as a seed verse that can bootstrap through "related verses" layer and bring in a network of other verses along with all Hadith and scholarly comments on these verses. Fig 2 depicts such a network of a concept generated from a single verse. In this manner, a single verse immediately leads to a network of verses, Hadith and scholarly comments thanks to the pre-annotated layers of related verses, Hadith and scholarly comments in our scheme taken from Tafsir Ibn Katheer.

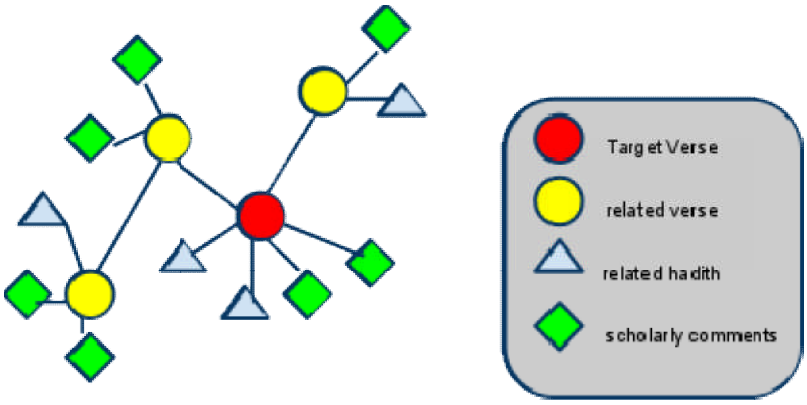


Fig 2. A conceptual cluster of related verses, Hadith and scholarly comments generated from a single verse

When no initial verse can be identified from user query, the other NLP layers can still lead to valuable information. User query and Qur'anic verses can be searched for matching named entity or matching semantic frames, and those potential verses can then be linked through the three Tafsir layers for more information. Our system should be able to deal with user input both in Arabic and English. Corpus layers and contents are mainly in Arabic, but as depicted in Fig 1, an English translation layer will act as an interface between user request in English and subsequent analysis and search for verses in Arabic.

Through this novel multi-layered extensive annotation of the Qur'an along with information from Hadith and Tafsir, it is likely that many queries can be answered about apparently contradicting verses from the Qur'an. For example, while verse 2:219 states that "They ask you about wine and gambling. Say, 'In them is great sin and [yet, some] benefits for people'" indicating permission, the verse 5:90 states "O you who have believed, indeed, intoxicants, gambling, ….are but defilement from the work of Satan, so avoid it.." which clearly indicating prohibition. When referencing Hadith and other scholarly comments, it would be evident that the later verse chronologically revealed at a later stage abrogating the initial permission expressed in the former verse, in order to prepare gradually a community of Muslims who are so used to consumption of alcohols.

Our pilot annotation of chapter 2 of the Qur'an will enable developing techniques to allow semi-automatic annotation of the other layers. This extensive corpus of the Qur'an and Tafsir is first of its kind and is likely to find many useful applications of knowledge extraction and other NLP tasks. Such application is expected to attract many audiences. This application will enable Qur'anic scholars, students as well as general Muslims to retrieve information from Tafsir. Also, it will enable Arabic linguists to search and analyze various linguistic features of the Qur'an. The extensive annotation guideline of this project might create interest among Arabic corpus linguistics community to produce specialized corpora in different genre. Also, this multi-layer annotation of classical Arabic text might

act as a training set for Arabic computational linguists and machine learning algorithms for automatic annotation of a rich library of classical Arabic texts. Also, this multi-layer annotation model can be replicated for similar set-up like the Bible and its commentaries.

REFERENCES

Ad-Dinawry, A. M. (died in 855) (1978). Tafsir Gareeb al-Qur'an. Dar al-Kutub al-Elmiyyah.

Al-Baghawi, A. M. (died in 1089) (1997). Ma'alim at-Tanzeel Dar Taibah lin-Nashr wat-Tawze'a, 4th Edition

Alsaif, A. and K. Markert. (2009) Arabic discourse annotation manual. University of Leeds.

As-Sabt, K. A., (1996) Qawaed at-Tafseer. PhD thesis, Islamic University of Medinah.

At-Tabari, M. J. (died in 889) (2000). Jeme' al-Bayan fi Ta'weel al-Qur'an. Moassat al-Risalah.

Burchardt, A., K. Erk, A. Frank, A. Kowalski, S. Pado, and M. Pinkal. (2009). Using framenet for semantic analysis of German: annotation, representation, and automation. To be published.

IbnKatheer, I. O., (died in 1353) (1999).Tafsir Al-Qur'an Al-Adhym. Dar Taibah lin-Nashr wat-Tawze'a.

Dror J., D. Shaharabani, R.Talmon and S. Wintner. (2004) Morphological Analysis of the Qur'an. *Literary and Linguistic Computing*, 19(4):431-452

Dukes, K and N. Habash (2010a) *The Crescent Corpus: Morphological Annotation of the Holy Qur'an*. Submitted to the seventh international conference on Language Resources and Evaluation (LREC-2010). Valletta, Malta.

Dukes,K. A. Sharaf and E. Atwell. *(2010b) Annotation Guidelines for the Classical Arabic Dependency Treebank*. Submitted to LREC-2010. Valletta, Malta.

Elkateb, S., Black. W., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C. (2006) Building a WordNet for Arabic, in Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.

Prasad, R. Dinesh, N., Lee,A., Miltsakaki,E., Robaldo, L., Joshi, A. and Webber, B. (2008). The Penn Discourse Treebank 2.0.Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Ruppenhofer,J., Ellsworth, M., Petruck,M. R. and JohnsonC. R. (2005). FrameNet: Theory and Practice.

Sharaf, A. and Atwell, E., (2009). A Corpus-based computational model for knowledge representation of the Qur'an. In the fifth corpus linguistics conference, Liverpool

Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, E., Zaenen, A., Rayot, S., and Antoniadis, G. (2000). Annotating a large corpus with anaphoric links. In Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000). Lancaster, UK

APPENDIX 2 – CL2009 PAPER

KNOWLEDGE REPRESENTATION OF THE QURAN THROUGH FRAME SEMANTICS

A corpus-based approach

Abdul-Baquee Sharaf *Eric S. Atwell*

School of Computing

University of Leeds

Leeds, LS2 9JT

United Kingdom

{scsams,eric}@comp.leeds.ac.uk

Abstract

In this paper, we present our in-progress research tasks for building lexical database of the verb valences in the Arabic Quran using FrameNet frames. We study the verbs in their context in the Quran, and compare that with matching frames and frame evoking verbs in the English FrameNet. We analyze the gaps and make appropriate amendments to the FrameNet by adding new frame elements and relations.

1. Introduction

The Quran is the central religious text of Islam – the world's second largest religion with a growing population of over 1.5 billion Muslims (1). Muslims believe that the Quran contains the words of God revealed on Prophet Muhammad by the Angel Gabriel (2); and that it is free from contradictions or discrepancies (3).

While there has been research in Arabic corpus linguistics (Atwell et al 2008) (Al-Sulaiti & Atwell 2006), or keyword search tools for the Quran (4), to our knowledge no extensive work has been done towards Quranic Corpus Linguistics. The goal of this work-in-progress research is to design a Knowledge Representation (KR) model for the Quran leveraging on the concept of 'frame semantics' as introduced by Fillmore (Fillmore 1978). Based on the concept of frame semantics, researchers in 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.

In our research project, we aim to build a FrameNet like lexicon for the verbs in the Arabic Quran. This initial attempt will enable future extension to include predicates other than verbs and to consider other classical Arabic texts as well as Modern Standard Arabic.

This paper is laid out as follows: Section 2 gives background information on Arabic verbs and some linguistic style of the Quran. Section 3 gives a sketch of related works on Quranic and Arabic verbs. Section 4 gives background information on the FrameNet lexicon. Section 4 details our intended research task and the challenges towards its implementation. Section 5 describes Framenet integration projects for other languages. Section 6 reports on the main tasks and challenges of this project. Finally we conclude highlighting the novelty of our research and its expected benefits.

2. Backgrounds

2.1 Arabic Verbs

In general, classical Arabic follows Verb-Subject-Object (VSO) order. The majority of Arabic verbs are trilateral, which can be derived to 15 different forms. Each derivation signifies some semantic variations over the original form. Table 1 gives a brief account on the most frequent nine such forms with their semantic significance. (Wright 1996) provides more elaborate discussion. The semantic significance of each derivation form is a subtle aspect of Arabic grammar which has no direct equivalent in the grammar/morphology of English or European languages.

NO	pattern	Semantic significance	Examples
I	فَعَلَ Fa3aLa	• When the 2nd radical is vowelized with (a) it mostly indicates transitive. • When the 2nd radical is vowelized with (i) it mostly indicates intransitive.	كَتَبَ to write فَرِحَ to be glad
II	فَعَّلَ Fa33aLa	• Intensive or extensive meaning of the first form • Convert the intr. In 1st form to transitive • Estimative or declarative	كسَر (to break) and كسَّر (break into pieces) فرح (to be glad) فرّح (to gladden) كذَب (to lie), كذّب (to call one a liar)
III	فاعَلَ Faa3aLa	• Place effort to perform act upon the obj. • Convert prepositional object to direct obj.	قاتله (he tried to kill him) كتب إلى (write to) = كاتب (write to) خاشنه (he treated him harshly)

		• Use Quality or state to affect another person	
IV	أَفْعَلَ aF3aLa	• Factitive or causative • Denominative (derive from noun a tr. Verb) • Movement towards a place/time	جلس (to sit down) and أجلس (to dib one sit down) أثمر (to bear fruit ثمر) أشأم (to go to Syria الشام) أصبح (to enter upon the time of morning الصباح)
V	تفَعّل taFa33aLa	• Express the state into which the obj. of the 2nd form was brought into action • Reflexive or effective	تكسّر (to be broken in pieces) علّم (to teach) and تعلّم (to become learned)
VI	تفاعَل taFaa3aLa	• Express the state into which the obj. of the 3rd form was brought into action • Convert the tr. Sense of 3rd form to reflexive • Reciprocity	باعدته (I kept him aloof) فتباعد (so he kept aloof) تماوت (to pretend to be dead) قاتله (he fought with him) and تقاتلا (the two fought with one another)
VII	انْفَعَل inFa3aLa	• Non-reciprocal but reflexive significance of the 1st form • A person allows an act to be done in reference with him	انكسر (to break [intr.], to be broken) انهزم (to let oneself be put to flight, to flee
VIII	افْتَعَل iFta3aLa	• Reflexive or middle voice of the 1st form. • Reciprocal	عرض (to place smth before one) and اعترض (to put oneself in the way, to oppose) اقتتل الناس (the people fought with one another
X	استَفْعَل istaF3aLa	Convert the factitive significance of the 4th form into the reflexive or middle A person thinks that the quality expressed in 1st form is applicable to himself A person seeking what is expressed by 1st form	أسلم (to give up) and استسلم (to give oneself up, to surrender) حلّ (to be lawful) and استحل (he thought that it was lawful for himself to do) غفر (to pardon)استغفر (to seek pardon)

Table 1. Most common forms of Arabic trilateral verbs.

2.2 The Quranic Linguistic Style

According to Muslims, the Quran is divine and contains words of God. It was revealed over a period of 23 years to the Prophet Mohammad in Arabic language. It contains around 78,000 words within the 114 chapters. The central topic of the Quran is to establish the monotheistic creed of God being the only possessor of divine power and only being who deserves to be worshiped. Prophet Muhammad challenged the Arabs to find another text –or a chapter of a text- like the Quran (5). The Quran claims to contain the fairest of statements and a scripture able to raise emotions and sentiments (6).

Following are some of the characteristics of the linguistic styles in the Quran. These features should pose special interests and challenges for computational linguistics solutions.

2.2.1 Scattered information on a same topic

The Quran often talks about a topic scattered within many different verses in different chapters. Consider the following verses (7):

[1] *Show us the straight path, The path of those whom Thou hast favoured* [1:6,7]
[2] *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* [4:69]
[3] *He who holdeth fast to Allah, he indeed is guided unto a right path* [2:101]

In [1] 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 [2] which is in a different chapter gives four types of people whom God shown favour. In [3], which is again in a different chapter, expands this list of favoured category to include one more.

The Quran also repeats a certain story, for example, of a previous prophet in many chapters but each occurrence adds certain information not present in other occurrences. For example, the Quran tells various aspects of the story of Moses in 132 places distributed among 20 chapters. This feature of the Quran makes a good case for computational solutions towards bringing these scattered occurrences automatically in one thread.

2.2.2 Literal vs. technical sense of a word

The Quran borrows an Arabic word and specializes it to indicate a technical term. Consider for example the word جَنَّة/jannah meaning literally 'a garden', but -as a technical term- in the Quran whenever this word is used it refers to 'the paradise' where the believers will abode as reward after the Day of Judgment. However, there are few instances where this word is used in the literal meaning to refer to certain gardens in this world. In the following examples [4] uses the more frequent technical sense and [5] uses the less frequent literal meaning.

[4] *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); [3:133]*
[5] *There was indeed a sign for Sheba in their dwelling-place: Two gardens on the right hand and the left..[34:15]*

2.2.3 Grammatical shift

The Quran often draws the attention of the reader by shifting grammatical agreement in a statement. For example, in [6] the mode changed from 'you' to 'they' and 'them' moving from 2nd person to third person. In [7] the verse shifted from addressing the Prophet alone to addressing the group.

[6] *when ye are in the ships and they sail with them with a fair breeze* [3:133]
[7] *O Prophet! When ye (men) put away women*..[65:1]

2.2.4 Verbs associating with different preposition

The Quran 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 [8a] and [8b] 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 Quran 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 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 Quran by using a verb (be alone) with a preposition (to) from another verb 'go' conveyed the meaning of 'being alone and going to' at the same time. This unique characteristic made both translations in [8a] and [8b] partially true, highlighting either the sense of the original verb 'be alone with' as in [8a] or the implicit verb with explicit preposition 'go to' as in [8b]. See Ibn-Katheer (2006) on his commentary of this verse.

[8a] *When they meet those who believe, they say: "We believe;" but when they are <u>alone with</u> their evil ones, they say: "We are really with you: We (were) only jesting."* [2:14 Yusuf Ali Translation]
[8b] *And when they fall in with those who believe, they say: We believe; but when they <u>go apart to</u> their devils they declare: Lo! we are with you; verily we did but mock.* [2:14 Pickthal Translation]

2.2.5 Metaphors and Figurative

The Quran uses a lot of metaphors and figurative language. In [9] Pickthal used the verb 'shine' but the Arabic verb /ishtala means 'to flare' and shows the analogy of 'old age symptom by many gray hair' with a 'fire burning a bush'. In [10] the Muslim army was so frightened that it felt as if their hearts reached to the throats.

,

[9] *My Lord! Lo! the bones of me wax feeble and my head is shining with grey hair..*[19:4]
[10] *When they came upon you from above you and from below you, and when eyes grew wild and hearts reached to the throats* [33:10]

2.2.6 Metonymy

In many verses the Quran uses metonymy. In [11] the Arabic verse literally means 'ask the town' which means (and was translated so) 'ask the people who live in the town'. In [12] 'a thing of planks and nails' is the 'Noah's ark', and in [13] 'eating food' metonymically means the 'need to answer call of nature', see Ibn-Katheer (2006) commenting on this verse.

[11] *Ask the township where we were, and the caravan with which we travelled hither.* [12:82]
[12] *And We carried him upon a thing of planks and nails* [54:13]
[13] *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 (earthly) food* [5:75]

2.2.7 Imperative vs. non-Imperatives

Arabic verbs are classified into past, present and imperative. Thus, in Arabic the imperative structure can be understood from the type of the verb used. However, in the Quran, although this general rule applies, yet there are many instances where imperative is understood although no imperative verb is used, for example in [14]. The opposite is also true: there are instances where an imperative verb is used, but the verse indicates non-imperative sense, for example [15] where the translator explicitly indicated the non-imperative meaning within brackets.

[14] *and whoever is minded to perform the pilgrimage therein there is no lewdness nor abuse nor angry conversation on the pilgrimage.* [2:197]
[15] *O ye who believe! Profane not Allah's monuments nor the Sacred Month nor the offerings nor the garlands, nor those repairing to the Sacred House, seeking the grace and pleasure of their Lord. But when ye have left the sacred territory, then go hunting (if ye will). [5:2]*

3. Related work

(Bielicky and Smarz 2008) describes building a valency lexicon for modern standard Arabic from the Prague Arabic Dependency Treeback (PADT). Their work is built on 'Functional Generative Description (FGD)' theory where verbs have valency frame with many complements known as functors which can further be divided into actants

(Actor, Addressee, Patient, Effect and Origin) and adjuncts (like Manner , Means and Location). This FGD concept was adapted for Arabic verbs and various corpus examples were drawn to prove the applicability of FGD for capturing Arabic verb valency. Some cases needed special attention like: diathesis, passive verbs, reflexivity and verb nominals.

(Al-Qahtani 2005) gives an extensive categorization of modern standard Arabic verb valence based on Case Grammar (CG) as described by (Fillmore 1968). Based on the assumption that CG is adequate to classify all verbs of a language and is universal across languages, Al-Qahtani went on to specify valence according to Cook's Matrix Model (Cook 1979) and its extension that includes 24 cells. According to this matrix five cases (Agent, Experiencer, Benefactive, Object, Locative) are plotted horizontally and type of verb (State, Process, Action) vertically. The date was taken from 8327 verbs from a lexicon (Al-Qahtani 2003) and most frequent 200 verbs were exhaustively sorted to a cell in the matrix, and thus proved the suitability of Cook's model for Arabic valence.

(Fiteih 1983) studied the prepositional verbs considering the Quran as his corpus. He could classify four classes of Quranic 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 [16]), or a nominal and a prepositional object (e.g., *send against someone something* as in [17]), or two prepositional objects (e.g., *come forth unto someone from some place* as in [18]), or one nominal object and two prepositional objects [19a] or one prepositional object and two nominal objects [19b].

[16] *And when he saw their hands **reached** not to it, he mistrusted them.. [11:70]*
[17] *For We **sent** against them a furious wind, [54:19 Yusuf Ali Translation]*
[18] *Then he **came forth** unto his people from the sanctuary [19:11]*
[19] *a. And Allah hath favoured some of you above others in provision [16:71]*
*b. He hath **bestowed** on those who strive a great reward above the sedentary[4:95]*

Shamsan (Shamsan 1986) studies the transitivity and intransitivity of Quranic verbs. He analyzed the valences of these verbs and tried to link between the form of these verbs and the semantic significance. He also observed the shift of a verb from intransitive to transitive sense based on semantic characteristics.

(Mir 1989) observed that quite a lot verbs in the Quran are used in idiomatic sense rather than literal meaning of the verb. He went on to list such expressions in the Quran. Some examples are given in the following quote.

> When a man's "eyes become cool", it means that he is pleased. A person who "brings down his wing" for you is being kind to you, but if he "bites his fingers" at you, he holds you a

severe grudge. If you think you lack the gift of fluent speech, you can pray to God to "untie the knot in your tongue" (Mir 1989: 2-3)

4. FrameNet Lexicon

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- FEs (BUYER, GOODS) and has many other non-core FEs (like: DURATION, MANNER, MEANS, MONEY, PLACE, PURPOSE, RATE, REASON, RECIPIENT, SELLER, TIME, UNIT).

Following are some illustrative examples from **commerce_buy** frame description. (The *lexical unit* is in **boldface** and *Frame Elements* are in CAPITALS).

[20] [BUYER Lee] **BOUGHT** [GOODS a textbook] [SELLER from Abby]
[21] Will they allow [BUYER you] to **PURCHASE** [MEANS by check?]
[22] [BUYER Sam] **BOUGHT** [GOODS the car] [MONEY for $12,000].
[23] [BUYER You] **BOUGHT** [RECIPIENT me] [GOODS three pairs] already!

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).

In addition to frame description, FrameNet also specifies frame-to-frame relations. These relations include: *inheritance, subframe, causative_of, inchoative_of* and *using*. For example, in figure 1, the frame **commerce_buy** inherits from more general **getting** frame, and is inherited by more specific **renting** frame, and is used by two related frames, namely, **importing** and **shopping**.

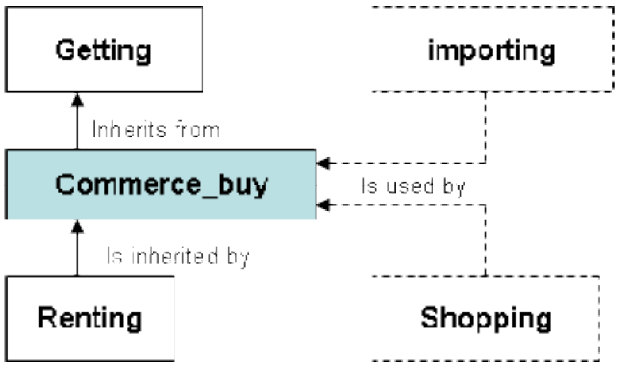


Figure 1. Frame-to-frame relations of the 'commerce_buy' frame

FrameNet also provides annotated sentences. This can be of two types: lexicographically motivated annotation and full-text annotation. In the former, the focus is to record the range of semantic and syntactic combinatory possibilities of a target lexical unit. Annotation of running text, on the other hand intends to exhaustively annotate each word in the text, which is possible thanks to layering techniques. The main layers are: a) Frame Element (FE) specifying frame elements as depicted in example [16] to [19], b) grammatical function (GF) like subject, object, etc., c) phrase type (PT) like noun phrase, verb phrase, etc and d) part-of-speech layer (POS).

Natural texts in many cases do not show up many conceptual frame elements. For this reason FrameNet annotation kept provision for 'Null Instantiation' (NI). This omission can be understood from the context and is called 'Definite Null Instantiation (DNI) like the missing RECIPIENT in [20] or cannot be retrieved but whose type is known like the missing QUARREL sense in [21], or the omission is allowed by the grammar of the sentence like the missing subject in any imperative structure like in [22].

[24] John contributed $20.
[25] Bob and Sue would argue all day.
[26] Get out immediately!

Since the launch of the English FrameNet, many researchers started to use FrameNet for various applications for example, *Machine Translation* (Boas 2002), *Question Answering* (Narayanan & Harabagiu 2004), *information retrieval* (Narayanan & Mohit 2003), *textual entailment* (Burchardt & Frank 2006), and also by incorporating it into domain specific ontology like BioFrameNet project (Dolbey et al. 2006).

5. Multi-lingual FrameNet projects

Since the release of the English FrameNet, researchers started similar projects in other languages. Successful examples are German, Spanish and Japanese.

5.1 German SALSA project

The German FrameNet project known as SALSA (Burchardt et al 2009) builds on the assumption that the English FrameNet is based on coarse-grained semantic classes which describes prototypical situations and thus, can be applied to other languages. During the course of the project, the team have found high correlation between English and German frames. However, they encountered some problems related to non-existence of certain language constructions in English (like some use of datives) and lexicalization differences in certain semantic domains (such as movement). The team went on to exhaustively annotate a large scale German corpus – the TIGER treebank (Bransts et al. 2002) – and in the process they had to encounter issues which were not faced by the FrameNet team, like dealing with idioms, support verb constructions, and metaphors. Idioms are multiword fixed expressions, and hence, the team decided to consider the whole expression as a frame-evoking word. In support verb constructions, the verb only supports a head noun (like 'give lecture') where the 'lecturing' frame should evoked instead of a 'giving' frame. The SALSA team in this case annotated the verbal part with a pseudo frame 'Support' with the noun as SUPPORTED frame element. In case of metaphors, in order to understand the literal source meaning should be transferred to the target intended meaning. The SALSA teams decided to annotate such cases with two frames: one for the target and one for the source.

As FrameNet is still under development, the team had to encounter non-existence of certain lemma senses in the English FrameNet. In these cases, they created proto-frames which define a new Frame following the style of the English FrameNet, and are also included in the frame-to-frame relationships.

The annotation is done using home-made SALTO tool that extends the TIGER syntactic tree to include Frame description. Unlike FrameNet, SALSA annotates frames with only 'core' frame elements.

5.2 Spanish FrameNet

Spanish FrameNet (Subirats & Petruck 2003) uses the English FrameNet lexicon to build a Spanish lexical resource. The project built a subcorpus of sentences from a 300 million word Spanish corpus that contains texts from various genres. (Subirats & Petruck 2003) report some difference in the lexicalization patterns of emotion predicates between English and Spanish as follows:

> "While both languages lexicalize the causative meaning with a verb (*sorprender* and *surprise*) and the stative meaning with an adjective (*estar sorprendido* and *to be surprised*), Spanish lexicalizes the inchoative meaning in the reflexive verb *sorprenderse* - 'to get surprised', while English uses a construction with *get* and the adjectival past participle *surprised*. In addition, while English has just one lexical unit *surprised* in the `Experiencer_subject` frame, Spanish has two: *sorprendido* used in conjunction with *estar* as a stative; and *sorprenderse* which is inchoative."

5.3 Japanese FrameNet

Japanese FrameNet (Ohara et al 2004) is a project started in 2002 based on English FrameNet. It started with a pilot study of motion and communication verbs. Corpus evidence is taken from the Mainichi newspaper corpus. The project team realized that unlike English, Japanese specifies a path along with motion, and thus has verb for 'go across' and another for 'go beyond, go over'. Therefore, they suggest amending Frame elements with BOUNDARY or ROUTE elements.

6. The Quranic FrameNet Project

6.1 Main Tasks

The first task is to collect all verbs in the Quran and their context in the verses. The reason we chose to consider only verbs is: first, to start with a feasible scope, and second, in Arabic –as well as other languages- verbs play the most vital predicate role. Malise Ruthven explains further:

> Substances and adjective are almost always verbal derivatives, usually participles or verbal nouns. A clerk is a writer [katib], a book is a writ [kitab]. Aeroplanes and birds are thing that fly [tiara and tayr]…it is precisely because Arabic refrains from classifying words into discrete particles, but keeps them instead in a logical and balanced relationship with a central concept. –the verbal root – that it becomes an eminently suitable language for religious expression." (Ruthven 1984:111)

This work of Quranic verbal analysis is being carried out through a machine readable index of the Quran (Abdulbaqi 1955). Each verb will be classified into their form (see verb forms in Table 1), which will help in semantic labeling later. Then, each Quranic verb needs to be studied to find a matching FrameNet lexical unit. For ambiguous cases, several parallel English translations will be consulted. Also, Books of Tafsir (scholarly interpretation of the Quran) for example (Ibn-Katheer 2006) or specialized lexicons and dictionaries (for example (Ibn-Mandhour 1997) or (Penrice 1873)) can be studied for clarification. Through this chosen lexical unit, the

corresponding frame in FrameNet will then be studied for appropriateness. To check this 'appropriateness', all target Quranic verb valences must exhibit the core frame elements of the chosen frame.

As an example, consider the *Ingestion* frame as depicted in Table 2 below. This frame has two core elements: ingestibles and an ingestor.

Frame Name	**Ingestion**
Definition	An Ingestor consumes food or drink (Ingestible), which entails putting the Ingestible in the mouth for delivery to the digestive system. This may include the use of an Instrument. Sentences that describe the provision of food to others are NOT included in this frame.
Core Frame Elements	**Ingestibles** The Ingestibles are the entities that are being consumed by the Ingestor. **Ingestor (Sentient)** The Ingestor is the person eating or drinking.
Lexical Units	breakfast.v, consume.v, devour.v, dine.v, down.v, drink.v, eat.v, feast.v, feed.v, gobble.v, gulp.n, gulp.v, guzzle.v, have.v, imbibe.v, ingest.v, lap.v, lunch.v, munch.v, nibble.v, nosh.v, nurse.v, put away.v, put back.v, quaff.v, sip.n, sip.v, slurp.n, slurp.v, snack.v, sup.v, swig.n, swig.v, swill.v, tuck.v

Table 2: FrameNet description of the frame: Ingestion

Next, consider the verb 'eat' in the Quran. It appeared –with derived forms- 100 times. Table 3 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	**devour**	all that ye have prepared for them	12:48
F	they	**eat**	into their bellies nothing else than fire	2:177
G		**Devourer**	of unlawful	5:42

Table 3. Few KWIC lines for <eat> from the Quran

These Quranic usages mandate us to extend the FrameNet to capture these non-ingestible and non-sentient uses. Thus, we suggest following the German SALSA strategy of creating a proto-frame for this special sense of 'eating money'.

As indicated in the previous section, the Quran contains many instances of verbal idioms. In such cases, again we follow the SALSA solution of considering the whole multi-word idiom as a frame-evoking predicate. Similarly, in case of metaphors, we intend to produce two annotations of such verses: one for the literal meaning and another to represent the metaphorical intended meaning.

In addition to exhaustively annotating the subcorpus of verses containing verbs in the Quran, we intend to choose as a case study, full-annotation of chapter 2 'Surah al-Baqarah' as a sample chapter from the Quran. This chapter portrays vibrant use of verbs since 97.5% of its 286 verses contain verbs (Suleiman 1997). We will carry on annotation in three layers, as is the FrameNet practice: Frame Elements layer, Grammatical Function Layer and Phrase Type Layer. In order to annotate Grammatical function, we will resort to reference books which exhaustively analyzed the grammatical function of each verse of each chapter, for example (Salih 1998), and populate the grammatical function layer. It should be noted that because of the vocalized form of the Quranic text, many ambiguities that appear otherwise in modern standard Arabic will not be faced. However, it is evident that many Quranic expressions result in more than one valid syntactic –and semantic- tree. For example, consider [27] which can refer simultaneously to two valid meanings [27a] and [27b] depending on where to pause.

[27] *This is the book no doubt in it a guidance for those conscious of Allah* [2:2]
[27a] This is the book no doubt in it. It is guidance for those conscious of Allah.
[27b] This is the book no doubt. In it a guidance for those conscious of Allah.

6.2 Representation

To represent the frames and lexical units, we will adhere to the structure of the FrameNet Database as detailed in (Baker et. al 2003). The result will be presented online in the FrameNet style, where color highlighting will help distinguish various frame elements.

6.3 Evaluation and Applications

Our annotated Quran chapters should represent the following three verses which are related semantically but scattered in various locations. Using FrameNet's *Giving* frame and extending it to capture the non-profit charitable sense of spend, the following labeling can be made.

[28] and [they $_{\text{DONOR}}$] **spend** out of [what We have provided for them $_{\text{THEME}}$] [2:3]
[29] and they ask you what [they $_{\text{DONOR}}$] should **spend**. Say, "[the excess $_{\text{DONATED\_AMOUNT}}$]". [2:219]
[30] and they ask you what they should spend. Say, "Whatever [you $_{\text{DONOR}}$] **spend** of [good $_{\text{DONATED\_AMOUNT}}$] is [for parents and relatives and orphans and the needy and the traveler $_{\text{RECIPIENT}}$]. [2:215]

While [28] talks about the theme of the donated money, [29] qualifies the type of this theme to be from the excess money that is left after spending on the necessary needs. However, [30] gives an answer to the same question as in [29], but specifies the recipient of this spend rather than the type or amount of the money.

Similarly, proceeding with annotation of the Quran governed by frame semantics might reveal interesting findings which might not be captured in books of Tafsir (scholarly interpretation of the Quran).

Annotating the Quran with frame semantics will facilitate efficient search beyond the existing keyword search. A Quranic researcher will be able to search semantic frames and semantic roles in addition to keywords.

Another interesting application of our semantically annotated corpus would be a Question Answering system. A question can be normalized into FrameNet style representation and matched with similar frames in the Quran for potential answers. (Shen and Lapata 2007) showed that FrameNet annotation produces significant improvement in QA systems.

6.3 Challenges

FrameNet is still under development. So for a certain lemma not all senses maybe covered. Also, because FrameNet only uses lexicographical prototype examples, some context usage might be hard to relate. Also, idioms and metaphors pose difficulty in representation. The lack of Arabic NLP tools –as compared to English NLP tools- might cause problems in automation and computational analysis.

7. Conclusions

We have embarked on a novel project towards frame semantics which starts by developing FrameNet frames for Quranic verbs, but can be extended to include

non-verbal predicated in the Quran and can further be extended to include predicates in Modern Standard Arabic. To our knowledge no previous attempts has been made towards integrating Arabic verbs to FrameNet frames.

Once completed, this research will benefit a wide audience. It will benefit Arabic NLP researchers considering a full-fledged Arabic FrameNet. It will benefit also the FrameNet community towards achieving a multi-lingual FrameNet project. This research will serve the wide Muslim population for better searching and extracting information from the Quran. In particular, the frame reports of Quranic Verb will interest Arabic linguists is analyzing the valence of the Quranic verbs.

NOTES

(1) http://www.adherents.com/Religion_By_Adherents.html

(2) The Quran 26: 192-195

> *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.* [Pickthal Translation]

(3) The Quran 4: 82

> *Will they not then ponder on the Qur'an? If it had been from other than Allah they would have found therein much incongruity.* [Pickthal Translation]

(4) See for example http://www.searchquran.org

(5) The Quran 10:38

> *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.* [Pickthal Translation]

(6) The Quran 39:23

> *Allah hath (now) revealed the fairest of statements, a Scripture who's parts resembling each other, paired whereat doth creep the flesh of those who fear their Lord, so that their flesh and their hearts soften to Allah's reminder*

(7) Citing verse reference with notion [x:y], x indicates chapter number and y indicates verse number. Unless otherwise mentioned, all translations are taken from Pickthal's translation available at University of Southern California's Centre for Muslim-Jewish engagement website: *http://www.usc.edu/schools/college/crcc/engagement/resources/texts/muslim/quran/*

REFERENCES

Abdulbaqi, M. (1955*) Aphabatical Index of the Quranic Words* (In Arabic). Dar al-adhami, Beirut. Available online

http://www.qurancomplex.com/IdIndex/default.asp?TabID=1&SubItemID=7&l=arb&SecOrder=1&SubSecOrder=7#

Al-Qahtani, D (2005) *Semantic Valence of Arabic Verbs*. Librairie du Liban Publishers

Al-Sulaiti, L. and E. Atwell (2006). "The design of a corpus of contemporary Arabic." *International Journal of Corpus Linguistics*. 11, 135-171.

Aston, G. and L. Burnard, (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press.

Atwell, E., N. Abbas, B. Abu-Shawar, A. Alsaif, L. Al-Sulaiti, A. Roberts and M. Sawalha (2008*).* "Mapping Middle Eastern and North African diasporas: Arabic corpus linguistics research at the University of Leeds" *In: Proceedings of BRISMES Conference 2008*.

Baker, C., C. Fillmore and B. Cronin (2003). "The Structure of FrameNet Database". *Int. Journal of Lexicography*, 16(3), 281-296.

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.

Bielicky, V and O. Smarz (2008) "Building the Valency Lexicon of Arabic Verbs", LREC 2008.

Boas, H. (2002). Bilingual framenet dictionaries for machine translation. In Proceedings of the 3rd International Conference on Language Resources and Evaluation. Spain.

Brants,S., S. Dipper, S. Hansen, W. Lezius and G. Smith (2002). The TIGER Treebank.

Burchardt, A. and A. Frank (2006). Approximating texual entailment with LFG and FrameNet frames. In Proceedings of the 2nd Recognising Texual Entailment Workshop. Venice, Italy.

Burchardt, A. K. Erk, A. Frank, A. Kowalski, S. Pado, and M. Pinkal(2009): *FrameNet for the semantic analysis of German: Annotation, representation and automation (preprint).* In Hans Boas (ed.): Multilingual FrameNet. Mouton de Guyter. (In Press)

Cook, W. (1979) Case Grammar: Development of the Matrix Model. Washington, D.C. : Georgetown University.

Dolbey, A. M. Ellsworth, and J Scheffczyk (2006). BioFrameNet: A domain-specific fragment extension with links to biomedical ontologies. In Proceedings of the Biomedical Ontology in Action.

Fiteih, M. (1983) Prepositions and Prepositional Verbs in Classical Arabic. PhD thesis, Univesity of Leeds.

Fillmore, C.(1968) The case for case. In Bach, E. W. and Harms, R.T. (Eds), *Universals in Linguistic Theory*.

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.

Ibn-Katheer (2006). *Tafseer Al-Quran* (In Arabic). Dar Al-Kutub al-Elmiyyah.

Ibn-Mandhour (1997) *Lisan Al-Arab* (In Arabic). Dar Sadir.

Mir, M (1989) Verbal Idioms of the Quran. Michigan Series on the Middle East, No. 1.Center for Near Eastern and North African Studies, University of Michigan, Ann Arbor.

Narayanan, S. and B. Mohit (2003). Semantic extraction with wide-coverage lexical resources. Companion Volume of the Proceedings of HLT-NAACL. Canada, 2003.

Narayanan, S. and S. Harabagiu (2004). Question answering based on semantic structures. In Proceeding of the 20th international conference on Computational Linguistics. Switzerland, 2004.

Ohara, K., S. Fujii, T. Ohori, R. Suzuki, H. Saito, and S. Ishizaki (2004). "The Japanese FrameNet Project: An introduction." LREC 2004. The Fourth international conference on Language Resources and Evaluation. Lisbon, Portugal. May, 2004.

Penrice, J. (1873). *Dictionary and Glossary of the Koran*. Adam Publishers and Distributors, India.

Ruppenhofer, J., M. Ellsworth, M. Petruck, and C. Johnson (2005). "FrameNet: Theory and Practice.

Salih, B. (1998). Al-E'rab al-Mufassal Li-Kitabillah al-Munajjal (Arabic). Dar Al-Kutub Al-Elmiyyah. Beirut.

Shamsan, A. (1986). Al-Fi'l fil Quran al-Kareem. (In Arabic). Kind Saud University.

Shen, D. and M. Lapata (2007). Using semantic roles to improve question answering. In Proceedings of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning, Prague, pp. 12–21.

Subirats, C.and M. Petruck, (2003). Surprise: Spanish FrameNet. *International Congress of Linguists. Workshop on Frame Semantics, Prague (Czech Republic), July 2003*.

Suleiman, F (1997). Al-Fial Fi Surat Al-Baqarah (In Arabic). Maktabat al-Aadab. Cairo

Wright, W. (1996) A Grammar of the Arabic Language. Librairie Du Liban, Beirut.

APPENDIX 3 – TRAINING AND DEVELOPMENT NEED ANALYSIS

University of Leeds **Date……**

PGR Training and Developmental Needs Analysis[1]

This audit is designed to be used by you to identify (from the Joint Skills Statement) where your personal strengths and weaknesses are. Use the statements on the following pages to provide a picture of which areas you are particularly confident and competent in and which require some work. You may wish to seek clarification from your supervisor about what individual elements of this audit mean within your discipline.

We recommend that you audit your own personal training and development needs at least once every six months. Record the date of the audit so you can chart how you have developed whilst a research student.

You can download copies of this document at http://www.leeds.ac.uk/rtd/

It is vital that you audit yourself with complete honesty. To help you, we suggest the following broad levels of competency and awareness:

Competency Levels:

1 = Good first degree candidate
2 = A research student with a little experience
3 = A more experienced PhD student
4 = A competent and confident doctoral candidate
5 = A truly outstanding doctoral candidate – able to teach, train or coach research colleagues in this area.

And finally, before you enter a score for any statement, think to yourself – how can I provide *evidence* for this?

[1] Adapted with thanks from the University of Manchester PG Development Needs Analysis

A. Research Skills and Techniques

Be able to demonstrate:	**More specifically:**	**Score**	How to Improve
A1. The ability to recognise and validate problems	Able to define original research problems		
	An understanding and application of appropriate research philosophies		
	Able to write a research proposal, to the level required of applications for postdoctoral work	**3**	Check the guidelines for writing good research proposals
A2. Original, independent and critical thinking, and the ability to develop theoretical concepts	Able to formulate hypotheses and/or research questions for the purposes of designing a doctoral research project		
	Able to provide new and innovative research ideas and strategies		
A3. A knowledge of recent advances within one's field and in related areas	Have in place a strategy for keeping up to date with the latest publications from own and closely-related research areas.	**1**	- Need to know the main source of information - Need to find best way to keep information updated - Consider attending workshops, subscribe to RSS feeds
	Confident in searching for information in a variety of bibliographic and virtual sources	**1**	- Enrol into courses offered by Library
	Can communicate knowledgeably and debate concepts about their wider research area with academic colleagues	**2**	- Be part of research groups and explore possibility of presenting ideas to them
	Confidently able to manage any collected information so it can be retrieved and cited appropriately	**2**	- Enrol into a course on "managing PhD information"
A4. An understanding of relevant research methodologies and techniques and their appropriate application within one's research field	An understanding of what constitutes "high quality" academic research within ones field	**1**	- Discuss with Supervisor, colleagues, etc. - Read the most famous works from journals, etc.
	Have in depth knowledge and understanding of appropriate research techniques and their application	**1**	- Get hold to some good text book on NLP, CL, etc. - Resort to good wikis, encyclopaedia, etc. - Enrol into basic modules like AI32
	Able to demonstrate objectivity and carry out unbiased research	**2**	- Ask supervisor - Check for seminars and workshops on this topic
	Able to discuss and prioritise a range of methodologies to address a research question		- Have a dynamic list of the most popular works - Find mechanism to be kept updated
A5. The ability to critically analyse and evaluate one's findings and those of others	Ability to objectively acknowledge weaknesses and assumptions in one's findings. Ability to apply the same objectivity to the work of others	**1**	- Attend many seminars and observe how experts discuss and criticize - Have my work circulated to colleagues for feedback
	Can write a literature review of publishable standard on the topic		
	Good understanding of appropriate methods for		- Need to know what are these "testing methods"

Competency Levels: 1 = Good first degree candidate, **2** = A research student with a little experience, **3** = A more experienced PhD student, **4** = A competent and confident doctoral candidate, **5** = A truly outstanding PhD candidate

	testing conjectures or tentative conclusions		
	Where appropriate, an excellent IT ability in data collection, analysis and presentation in an appropriate graphical form		- Attend some presentation skills workshops
A6. An ability to summarise, document, report and reflect on progress	Able to objectively criticise own research and define future work		
	Able to maintain and use a research log or record of research activity commensurate with best practice from your discipline		
	Able to produce written summaries and progress reports of a variety of lengths to suit the purpose, and to an appropriate professional standard		
	Able to take regular reviews of own work to determine that it is of sufficient originality, quality and quantity to merit the award of a doctorate.		- find ways to present and sent own work for review within and outside Leeds

B. Research Environment

Be able to demonstrate:	More specifically:	score	How to improve
B1. Show a broad understanding of the context, at the national and international level, in which research takes place	Able to explain how research in own particular field is organised nationally in terms of (where appropriate) institutions and centres, congresses, societies, publications and some understanding of these internationally.	**1**	- ask supervisor - search for national and international research institutes - look for the renowned figures in the field
B2. Demonstrate awareness of issues relating to the rights of other researchers, of research subjects, and of others who may be affected by the research, e.g. confidentiality, ethical issues, attribution, copyright, malpractice, ownership of data and the requirements of the Data Protection Act	Be aware of the guidance offered to researchers at a national level (appropriate to your discipline) i.e. through RCUK, the OST, the NHS and relevant professional bodies etc. concerning ethical issues and ethical research practice within your discipline	**1**	- search for suitable workshops - search for university documents -
	Be fully aware of the University of Leeds' rules and regulations relating to academic misconduct (and particularly plagiarism)	**1**	- search University sites
	Be aware of University guidelines on intellectual property, copyright and ownership of research	**1**	- read those guidelines
B3. Demonstrate appreciation of standards of good research practice in their institution and/or discipline	A complete understanding of any relevant University guidelines on research practice (e.g. ethical practice) and any statutory regulatory requirements in your subject area	**1**	- search other universities pages
B4. Understand relevant health and safety issues and demonstrate responsible working practices	Be competent and competent in working with all relevant health and safety regulations	**1**	- already enrolled in Health and Safety Course
B5. Understand the processes for funding and evaluation of research	A broad understanding of how research is funded within ones own discipline and the mechanisms by which funding might be sought to continue ones current research	**1**	- search web
	Knowledge of how large and small-scale research proposals within your discipline are evaluated	**1**	- search web
B6. Justify the principles and experimental techniques used in one's own research	Have good knowledge of competing techniques and approaches in subject area and their relative strengths and weaknesses	**1**	- attend NLP courses
	Be able to justify and defend the decisions that underpin your research direction and methods		

Competency Levels: 1 = Good first degree candidate, **2** = A research student with a little experience, **3** = A more experienced PhD student, **4** = A competent and confident doctoral candidate, **5** = A truly outstanding PhD candidate

B7. Understand the process of academic or commercial exploitation of research results	Understanding of both procedures for submission and evaluation of research by journals and publishers and be able to prepare research results for submission.	**1**	- search
	Understanding of the major conferences in the research area.	**1**	- search
	Awareness of the various University facilities and support for exploitation of research.	**1**	- search

C. Research Management

Be able to demonstrate:	More specifically:	score	How to improve
C1. Apply effective project management through the setting of research goals, intermediate milestones and prioritisation of activities	Able to plan, organise and evaluate a research programme	**2**	Enrol in time & project magmt workshops
	Able to execute a research programme which follows a designated schedule to produce a finished thesis within the funded period	**2**	Enrol in time & project magmt workshops
	Able to set and prioritise a number of intermediate goals within an individual research project and to develop an effective strategy and timetable for meeting them	**2**	Enrol in time & project magmt workshops
	Able to make plans and balance competing demands on time effectively	**2**	Enrol in time & project magmt workshops
C2. Design and execute systems for the acquisition and collation of information through the effective use of appropriate resources and equipment	Able to collect and record information in an organised and professional way	**2**	Enrolled in a course “manage your PhD information”
	Competence in relevant data-collection and analysis software	**2**	Enrolled in a course “finding PhD information” and “speed PhD”
	Able to conduct searches using appropriate online and offline resources	**3**	Search resources
C3. Identify and access appropriate bibliographical resources, archives, and other sources of relevant information	Able to demonstrate an excellent awareness of potential sources of relevant information for subject area	**2**	
	Fluent in referencing appropriate sources (books, articles, websites, interviews and quotations) and able to use a variety of referencing styles and systems	**2**	
C4. Use information technology appropriately for database management, recording and presenting information	Able to establish a bibliography at the level expected for scholarly publication and keep it up to date through searches and electronic services	**2**	
	Able to use appropriate software to prepare extensive documents with any relevant special features, such as use of master documents and templates or embedding of charts, figures and	**2**	

Competency Levels: 1 = Good first degree candidate, **2** = A research student with a little experience, **3** = A more experienced PhD student, **4** = A competent and confident doctoral candidate, **5** = A truly outstanding PhD candidate

	images		

D. Personal Effectiveness

Be able to demonstrate:	More specifically:	score	How to improve
D1. Demonstrate a willingness and ability to learn and acquire knowledge	Able to identify and exploit sources of information or instruction on a new area.	**2**	
	Fully committed to, and engaged in, undertaking a meaningful research-specific and transferable skills training programme	**3**	
	Excellent attendance at seminars, meetings, workshops and conferences, evidenced with an up-to-date training record	**4**	
D2. Be creative, innovative and original in one's approach to research	Ability to generate new ideas and approaches	**3**	
	Ability to develop new methodologies as required	**2**	
	Ability to find and implement solutions to difficult problems	**2**	
D3. Demonstrate flexibility and open-mindedness	Able to analyse the strengths and weaknesses of one's own approach, and willing to complement it by an engagement with other approaches	**2**	**Through peer review and supervisor guidance**
	Be fully aware of all of the means of exploiting intellectual property and have considered the scope of knowledge transfer and entrepreneurial activity in relation to your research work	**1**	
D4. Demonstrate self-awareness and the ability to identify own training needs	Able to evaluate a wide range of skills, evaluate training needs in the light of this and the requirements of the research project, develop a coherent plan for future training	**3**	
D5. Demonstrate self-discipline, motivation, and thoroughness	Able to work to a professional level without supervision	**2**	
	Able to demonstrate high levels of accuracy, organisation and attention to detail commensurate with that of a professional independent researcher of your discipline	**2**	
D6. Recognise boundaries and draw upon/use sources of support as appropriate	Be able to objectively consider gaps in knowledge, understanding or ability and be aware of possible sources of support such as the skills of colleagues	**2**	
D7. Show initiative, work independently and be self-	Able to make and execute substantial research plans with guidance necessary only for specialist issues	**2**	

Competency Levels: 1 = Good first degree candidate, **2** = A research student with a little experience, **3** = A more experienced PhD student, **4** = A competent and confident doctoral candidate, **5** = A truly outstanding PhD candidate

reliant	Provide evidence of “academic independence” to colleagues and peers		

E. Communication Skills

Be able to demonstrate:	More specifically:		Your competence
E1. Write clearly and in a style appropriate to purpose, e.g. progress reports, published documents, thesis	Able to produce a well-structured and well written report of substantial length. Able to write concise, academic prose and express ideas with suitable clarity	**3**	
	Full mastery and control when writing a variety types of research document and in a variety of written styles	**2**	
E2. Construct coherent arguments and articulate ideas clearly to a range of audiences, formally and informally through a variety of techniques	Able to communicate own research orally, with proficiency and confidence	**3**	
	Able to explain own research at a range of levels	**3**	
	Able to produce well constructed clear presentations and use audiovisual aids where appropriate (slides, OHPs , PowerPoint)	**3**	
	Able to provide feedback around own research subject of the kind expected in referee's reports for journals and publishers and to respond to such feedback	**2**	
E3. Constructively defend research outcomes at seminars and viva examination	Able to present academic work at seminars and conferences fluently and confidently, and able to respond clearly and persuasively to questions and comments at such occasions	**2**	
	Confidently able to defend own work in meetings, at transfer, during academic interviews and during the viva	**2**	
E4. Contribute to promoting the public understanding of one's research field	Able to explain the importance and benefits of communicating your research outside of academia and understand how such activity fits with University and national policy direction	**2**	
	Able to write and present research in an appropriate manner for specialist or lay audiences, and be understood	**3**	

Competency Levels: 1 = Good first degree candidate, **2** = A research student with a little experience, **3** = A more experienced PhD student, **4** = A competent and confident doctoral candidate, **5** = A truly outstanding PhD candidate

E5. Effectively support the learning of others when involved in teaching, mentoring or demonstrating activities	Demonstrate an ability to effectively facilitate the learning of others and an ability to impart information effectively	**3**	
	Have an understanding a range of appropriate techniques for supporting the learning of others	**2**	

F. Networking and Teamworking

Be able to demonstrate:	More specifically:		Your competence
F1. Develop and maintain co-operative networks and working relationships with supervisors, colleagues and peers, within the institution and the wider research community	Regular attendance at conferences and meetings, awareness of other *researchers* in own and related fields	**2**	**Need to know the name of key players in the field**
	Both build and maintain co-operative networks and working relationships with supervisor(s), colleagues and peers within the University	**3**	**Participate in seminars and groups**
	Both build and maintain co-operative networks and working relationships with colleagues and peers in the wider research community	**2**	**Participate in conferences**
	Aware of and subscriber to appropriate virtual networks and sources of support (such as the University of Leeds RSU nets , ResearchResearch, UK GRAD and virtual subject specific networks such as JISCmail)		
F2. Understand one's behaviours and impact on others when working in and contributing to the success of formal and informal teams	Aware of the impact that own behaviours and actions have when building a healthy working relationship with supervisor(s)		
	I understand my behaviour and impact on others when working in and contributing to the success of formal and informal teams		
	Can work in teams (both inside and outside of academia) on often complex projects and can both reflect on quality of teamwork and solve team-working problems as they arise		
	Am aware of all the stakeholders of ones work, and have considered and acted-upon the best ways for interacting with them	**2**	**Need to know the role of Tutor**
F3. Listen, give and receive feedback and respond perceptively to others	Aware of techniques of giving and receiving feedback effectively		
	I am able to listen, give and receive feedback and respond perceptively to others		

G. Career Management

Be able to demonstrate:	More specifically:		Your competence (where 4 is an experienced PhD student)
G1. Appreciate the need for and show commitment to continued professional development	Active member of an appropriate professional institution or body		
	Regularly attend any appropriate departmental, school, faculty or University seminars and research meetings		
	Take some role in facilitating or organising seminars and research meetings, or some other form of administrative responsibility		
	Have ownership of, update and regularly review a personal development plan		
G2. Take ownership for and manage one's career progression, set realistic and achievable career goals, and identify and develop ways to improve employability	Be aware of potential employers, general recruitment practices and effective job hunting techniques		
	Have considered own career direction post-PhD and set realistic and achievable career goals		
	Have identified ways to improve my employability and acted upon them		
G3. Demonstrate an insight into the transferable nature of research skills to other work environments and the range of career opportunities within and outside academia	Be aware of the range of career opportunities within and outside academia		
	Be fully able to demonstrate the transferable nature of research skills to other work environments		
	Be aware of potential career paths stemming from the generic aspects of a PhD, including research techniques, project planning and communication skills		
G4. Present one's skills, personal attributes and experiences through effective CVs, applications and	A broad knowledge of types of CV's, interview techniques and standard questions and recruitment techniques such as psychometric testing.		
	Able to create a targeted CV which effectively presents own skills, attributes and experiences		

interviews	Able to present own skills, attributes and experiences effectively in a job interview situation		

Adapted with thanks from the University of Manchester PG Development Needs Analysis