Freie Universität Berlin

Fachbereich Geschichts- und Kulturwissenschaften

Hausarbeit

im Studiengang: „MA Arabistik“

Modul: Koran, Tafsir, Hadith

Thema:

Digitization of the Muṣḥaf

Gutachter: **Prof. Dr Islam Dayeh**

vorgelegt von:

Mahmoud Kozae

Berlin, 30.08.2019

Table of Contents:

1 Foreword

This paper investigates the current presence of Quran in digital media. The choice of the term *Muṣḥaf* for the title instead of Quran is in observation with the Islamic theological principle, which dictates that the Quran is an abstract word of God, while *Muṣḥafs* are material records of it, that were created by the Muslims to aid its recitation and preservation. This distinction is important in emphasizing the oral nature of the Quran and paying attention to the fact that any records will remain deficient, and cannot stand alone as a faithful depiction of the Quran without supplemental knowledge in the Islamic tradition on the guidelines, rules, and arts of reciting the Quran. However, both terms Muṣḥaf and Quran will be used interchangeably throughout the paper, as it is still common in scholarly circles and Muslim communities.

As the Quran is essentially an oral tradition, sound records seem to be more efficient in preserving and presenting it; and hence it can be argued that working on advancing techniques and methods for better curation and processing of sound data should be the sole focus in digitalizing the Muṣḥaf. However, this paper deals extensively with digital texts, as this form of data serves important functions in curating and advancing the software that works with sound records; these functions will be discussed in details in chapters 3 through 5.

2 The Quran in Print and Early Electronic Media

The first stage in modernity was the industrialization era, that dramatically reshaped human lives. When it comes to the lives of texts and their scholarship, the biggest effects came from the advances in printing, radio, and the phonographic records and its successor technologies. These technologies also affected the way various groups of Muslims access the Quran and interact with it. Nonetheless, there has always been a relatively wide time gap between the appearance of specific media technology and its adoption for Quran related affairs[1]. The factors for these delays were either certain counterproductive religious attitudes or simple lack of means and finances. The mentioned attitudes are the cautious fundamentalist ones regarding the involvement of any novelties in religious matters. As adoption of recent creations that did not exist in the Muslim tradition sometimes may constitute "unlawful innovations, *bidaᶜ* s. *bidᶜa*" that violate Islamic law. Such bans usually last until enough individuals of religious authority develop a sufficient understanding of the technology in question and start allowing its use, or until reality and the circumstance basically enforces a transition into the new technology without the need for explicit fatwas.

Nevertheless, it is hardly fair to entirely attribute the late adoption of modern technologies to religious dogma or even to consider it the principal factor. Any new technology needs enough time to prove its robustness and sustainability, and in a matter as sensitive as the curation of religious scripture, the excessive caution is entirely understandable. The proof of efficiency would then serve as justification and motivation for further financing and human engagement. Furthermore, scholars and religious authorities always presented elaborate reasoning and examination of the social changes and full context associated with adopting the new technologies, and they did not pronounce just simplistic dogmatic refutals; for example, the argument that the widespread of printed Quran might lead to its mishandling by non-muslims, or that running phonograph recitations in public locations would not be appropriate settings for listing to the holy text with the attention it deserves or raising concerns about the "purity" of radio stations and its association with entertainment; Which are all legitimate arguments from a religious standpoint[2].

In looking at the history of printed Muṣḥafs[3], a distinction has to be made by the first prints and the first widespread printed versions; there have been some attempts as early as the mid-1500s. However, no printed version became commonly used until the 1800s; principally among orientalists and Muslims in the Indian subcontinent. The widespread in

[1] Hirschkind (2003) p.342.

[2] Ibid. pp. 343-345

[3] Ibid. p.342.

the Arab world did not happen until the 1920s. The popularity of the printed Quran added many conveniences to the lives of Muslim scholars and people, but it can be argued that this transition did not radically affect the tradition of how the Quran has been transmitted and curated as much as it affected everyday lives of Muslims. The role of the printed Muṣḥaf merely assumed the role of the written Muṣḥaf as an aid-memoir for reciters and readers.

As soon as radio became common in Muslim countries, live broadcast of Quran recitation became a regular practice. Radio performances did not represent a change to reciters and resembled performances in mosques or other gatherings. In Egypt, a dedicated station *iḍāᶜat al-qurʾān al-karīm* for Quranic recitations was launched in the early 1960s[4]. It took longer for the Quran to be widely available in the earliest form of recorded sound media, the phonograph records. This technology had existed since the 1880s, and the first recorded recitations of the entire Quran became available in the mid-1960s. The collection of Quran recitations in the form of phonograph records was by a remarkable initiative by Labīb Al-Saᶜīd (d. 1988), who extensively described the project in a book published in 1969. An essential innovation of the project is the producing forms of recitations where reciters perform more slowly and with more emphasis on enunciations than in live recitations; this way the recorded recitation would be more suited for educational purposes[5].

The biggest shift that industrialist media caused is, that the reading of *Ḥafṣ ᶜan ᶜĀṣim* became dominant in large parts of the Muslim world[6].

[4] From the official website of the station *ertu.org*; accessed September 2019

[5] Al-Saᶜīd (1969) pp. 97-112.

[6] Hirschkind (2003) p.348.

3 Outline of Digitization and Computing

Digitization does not have a unified definition across domains, may it corporate, academic, or even private. Radical differences exist within each of the scopes mentioned above. A distinction can be made between the "digitization of resources;" which can be labeled as digital preservation or storage and the "digitization of activities" i.e. computing or computational support; which is simply the involvement of computers in human activities, or the complete delegation of specific activities to computers and machines i.e. full automation of tasks. Noteworthy is that digitalization is a human activity that is supported by machines to varying degrees from partial to full automation[7].

The definition of digitization is thus dependent on the intended form the produced digital data, which is determined by the forms of computing expected to be performed using it. For example, making image scans of books is practically digitization; however, it is an inadequate form of digitization if the intention is performing a full-text search or advanced text-analytics; a conversion of these image scans to digital text is then necessary. In the subsequent sections, elements of digitization and computing are discussed in detail while presenting their relations to Quran associated matters.

3.1 Data

In this day and age, data remains an ambiguous term. In both Oxford and Webster dictionaries, the first definition is that data is factual information and further definitions present as a term for information in digital form stored by a computer[8]. In practice, however, associating data with information has been shown to be imprecise[9]; as data by itself does not inform or further the knowledge of someone viewing it without supplementary analysis and elaborations, i.e., only the analysis turns data into information. Additionally, some collections of data might be highly unorganized and complex and would need years of investigation before it is possible to retrieve knowledge from it. Hence the act of collecting data would not necessarily involve or induce learning.

A heuristic way to look at data is to consider it records of observations, experiences, or measurements, that is retrievable a way that is sufficiently faithful to the original way it was observed, experienced, or measured[10]. The phenomenon of high-fidelity records was

[7] Bountouri (2017) pp. 29-50.

[8] Definitions retrieved September 2019 from the online portals *merriam-webster.com* and *lexico.com*

[9] Own view.

[10] Own definition.

not possible until modern times and the invention of electronic devices. Hence, people - and dictionaries- became accustomed to associate data with computers, which in a way is not wrong, but still reductive. Observations recorded on paper is practically data, that is not digital and is of low fidelity. Such case of non-digital data can be found in paintings of scenery that would never be as identical to the actual scenery as photographs. This distinction is necessary; for example, in the case of textual scholarship, manuscripts of interest are data that has not been digitized yet. Using the term data and characterizing which state each piece of data is essential, as such characterization is the first step in determining the further steps needed to process this data to extract meaningful results from it.

Digital data is the most versatile form of data and recording information in a digital format offers many advantages regarding its processing, transmission, and sustainable storage[11]. The predecessor to digital media was analog media, the comparison between the two is purely technical and is not relevant to the current discussion, it is enough to mention that the digital media offered many improvements over the analog ones in cost and quality-related matters.

When it comes to scholarship, a vital question is about determining which data is most relevant. Data collection and storage has never been more convenient and raises further issues on how to organize and curate data repositories. The data is also often multi-faceted, and in textual scholarship, there are almost always at least image and text data, and in case of Quranic studies, audio data is essential as well. Moreover, there are also the so-called meta-data which are mostly indexes and annotations to enable better organization and searchability of the data in question[12]. There are some standards on how to produce such meta-data but it is not always inclusive of all possible research questions and would not help every scholar in extracting the knowledge relevant a chosen research question. For example, a digital text corpus with extensive morphological annotations would not support a scholar examining figurative language in such a corpus. Always new methods on how to annotate data corpora are invented to enable more elaborate analysis and knowledge extraction.

Repositories of Quran data are abundant; this can be categorized as follows:

- Scans of Muṣḥaf manuscripts and their digitizations. Noteworthy are the collections of BnF and the Corpus Coranicum.
- Collections of texts. The most prominent example is the repositories of Tanzil project[13]; there exist many other collections in diverse formats adjusted for different purposes.

[11] Bountouri (2017) p. 37.

[12] Ibid p.32.

[13] More information on these collections in chapter 3.

- Annotated Corpora. The purpose of these corpora is to facilitate advanced analysis of the morphology and syntax. Most widely used are the Leeds Quranic Corpus and the Oujda Muṣḥaf Corpus[14].
- Libraries of recitations' sound records. These are very common around the web, usually, in MP3 format, the main difference among those libraries is the level of indexing; commonly they are retrievable per Sūra: one sound file for each Sūra, as in *islamway.com*. An important project is "Every Ayah Quran Files," which provides its collection as separate audio file per reciter and Ayah. Another innovative project is "Quran Word by Word," which provides a separate audio file for each word[15].

3.2 Theories and Methods

The first level of these theories is the mathematical ones; advancements in mathematics, while are not usually observable to the end-users, are always the basis for more efficient software. Other "low levels" are the advancements in manufacturing hardware and optimizations to programming languages; which all has an effect on digitization and computing activities yet are not of direct concern to a practitioner in e.g., humanities research.

Some higher-level methods[16] that are of direct relevance to Quranic research:

- Natural language processing: Which are different techniques to make human languages more understandable to machines; by allowing machines to read, e.g., morphology and sentiment.
- Machine Learning and Artifical Intelligence: sets of methods that make it possible for the computer to be better able in classifying data and predicting trends based on data.
- Speech feature extraction: turning sound data into mathematical representations; which is necessary to perform the previously mentioned NLP, ML, and AI on them.

Both methods for Arabic NLP and Speech Analytics are underdeveloped in comparison to English and many Latin-script languages; however, promising progress has been taking place in recent years[17]. There are many papers that present theories on how to better

[14] More information on these two corpora in chapter 3.

[15] More information on digital sound collections in chapter 4.

[16] This is a partial list; the development of software and computational techniques is backed by a large number of disciplines. Definitions are from Jurafsky and Martin (2009).

[17] Guellil et. al. (2019).

perform automatic speech analysis of Quranic recitations; however, most of them have not been utilized yet[18].

3.3 Applications

Digital Application is the utilization of methods to achieve specific results. The goals of implementing applications are either automating and accelerating task or accomplishing goals that human abilities and senses are not able to independently complete. It is important to differentiate applications from interactive software and platforms; as the specifications for a software usable by non-experts is a bit more extensive than an application; some examples of Quran related applications:

- AlQuran Cloud API[19]; which constitutes a helper for developers who need to display Quranic text in their applications. It provides an easy mechanism for automatically searching and fetching parts of the Quran.
- TextMiningTheQuran[20]; which are a series of studies based on the Leeds Quran Corpus that utilizes various NLP and statistical methods to characterize the language of the Quran.
- Tajweed Classifier[21]; an application of Machine Learning that based on Uthmanic text extrapolates where specific tajweed rules should be applied.

3.4 User Interfaces

User interfaces are parts of software that emphasize ease of use and access for non-experts. As many software and applications prerequisite advanced knowledge of underlying technologies. There are thousands[22] of Quran software with graphical user interfaces across many technologies and devices: PCs, tablets, and smartphones. It is a matter worthy of investigation, to tally which software is most widely used and which ones provide the most innovative features. However, there have been very few studies in that regard. Quranic

[18] This observation is made from examining multiple papers that propose computational methods for the Quran without providing access to software based on these methods. Some articles are discussed in sections 4.2 and 5.2

[19] *alquran.cloud/api*

[20] *textminingthequran.com*

[21] *github.com/cpfair/quran-tajweed*

[22] No extensive studies or statistics on the exact number of software exist; however, such a figure is highly plausible based on the number of results by searching for Quran software using a web search engine like Google, or application store for e.g., smartphone operating systems like Android or Apple IOS.

digital user interfaces provide a wide range of features, of reading the Arabic text of the Quran while being able to listen to recitations, and read tafsir or translations.

4 Muṣḥaf Digital Text

There are multiple research disciplines specialized in working with textual data; Computational and Corpus Linguistics besides Natural Language Processing. These disciplines deal with speech as well but to a much lesser extent, and dealing with the text still the main focus, as analyzing speech usually involves first generating text based on the speech then analyzing this text[23]. However, the processing of Arabic is still many steps behind Latin-script languages[24]. The first reason for this underdevelopment is that scholars mainly focused on replicating and reusing the techniques used for the other languages in processing Arabic, which were not always suitable for the nature of Arabic and ended up being less capable than intended. Another reason is that Arabic is not really one thing and strict distinctions between Classical Arabic, Modern Standard Arabic and the many Arabic dialects are not easily identifiable in a computational way i.e.; there are not enough digital data to enable automatic identification of the differences.

An example of the inefficiency of reusing the techniques of Latin-script languages is lemmatization, which is a standard step in the processing of natural languages. Lemmatization is assigning lemmas to words i.e., identifying to which lexical entry a word belongs. The roots of Arabic words have been used as lemmas in these lemmatization methods, which is not efficient as each root practically constitutes a chapter in an Arabic lexicon that might contain hundreds of entries. Assigning roots defies the purpose of lemmatization, which as disambiguation, as reverting each Arabic word to its root actually adds ambiguity.

In the following sections, some sources of Muṣḥaf digital text will be discussed, and some methods and applications that are geared towards processing texts.

4.1 Text Corpora and Repostories

4.1.1 Tanzil Project[25]

The encoding of the Arabic script went through multiple standards before the Unicode standard became stable and less susceptible to frequent updates and developments[26]. By the time when the Unicode became stable, there existed many corpora of Arabic texts in

[23] Jurafsky and Martin (2009) p. 54.

[24] Guellil et. al. (2019).

[25] The information is from: *tanzil.net/docs/tanzil_project*, accessed September 2019.

[26] Guellil et. al. (2019).

various inhomogeneous encoding standards, among which corpora of Quranic text. In addition to non-uniformity of encoding, there were many erroneous texts that one would stumble upon while searching for the Quran using search engines, mainly Google. Tanzil Project, which was a community effort and not carried out by a scholarly, religious, or political institute, successfully complied a uniformly encoded error-free digital corpus of Muṣḥaf text. The corpus has never been subjected to rigorous scholarly verification; however, it seems to have become the de-facto standard[27] digital Muṣḥaf, and hundreds of websites, applications, and even scholarly projects that work with digital Quran text, cite the Tanzil Project as their source for this digital corpus.

4.1.2 Oujda Al-Muṣḥaf Corpus[28]

This corpus provides a fully word-by-word annotated Muṣḥaf. The annotation is concerned with the morphology and involves six categories:

- Stem: the main host morpheme of a word without suffix or a prefix.
- Lemma: for verbs, lemmas are the third person masculine singular perfective, and for nouns, lemmas are the singular form.
- Root: the conventional Arabic root.
- Part of speech: the type of the word, for which there are more than 100 possibilities
- Stem pattern: the *wazn* of the stem.
- Lemma Pattern: the *wazn* of the lemma.

The corpus is an impressive achievement, and a preliminary examination of it shows no mistakes. However, a more thorough examination is still needed. The authors have also presented a seemingly practical solution for lemmatizing, the effectiveness of this solution needs further testing as well. An excerpt from the corpus is in the appendix page 18.

4.1.3 Leeds Quranic Arabic Corpus[29]

This project aimed at more sophisticated forms of annotation. In addition to morphology, it contains layers of description for syntax and explanation of word dependency; all of which are presented in a form of well-designed visualizations with extensive explanations. The corpus has been an important resource for computer-assisted Quranic Studies and has been utilized and referred to in tens of papers. An example of how a verse is annotated in this corpus is in the appendix page 19.

---

[27] A list of websites that use this project is in *tanzil.net/docs/who_is_using_tanzil*; most of which are widely used.

[28] Zeroual and Lakhouaja (2016).

[29] Dukes (2013) pp. 141-179.

4.2 Applications and Methods Based on Textual Data

4.2.1 Tajweed Classifier[30]

This is another community-driven project and not a product of scholarly research. This project implements an automatic detection and annotation of the rules of *taǧwīd* based on the Uthemanic Rasm. It utilizes decision trees to process the text letter by letter, applying elaborate algorithms to check whether the position is associated with a certain *taǧwīd* rule.

4.2.2 Integrity and Authenticity Checking

Alsamdi and Zarour (2015), presented a method to automatically verify if a digital text that is claimed to be from the Quran is indeed so. The method is independent of the vowel marks *ḥarkāt* and the additional writing signs *taškīl*; however, it also verifies the correctness of those marks and signs if they are present. The method utilizes a computational hashing technique that associates each verse with a numerical representation, called "hash", and to verify the authenticity of a specific digital text, it is converted to a hash as well, and the two hashes are then compared. This method allows for reliable, fast authenticity checks.

4.2.3 Extracting Semantic Relations

Extracting semantic relations is a process in Natural Language processing that has been underdeveloped in Arabic. Having such software allows for advanced automated text analytic, like finding entity names within a text, finding words that have specific relations to one another like synonyms, antonyms, meronyms, etc. Bentrica, Zidat, and Marir (2017) presented a method to achieve this type of automated processing using the Leeds Quranic Arabic Corpus. The technique had a precision rating of 84%; the scholars attribute the deficiency partially to mistakes in the annotations of the used corpus.

[30] Project website*: github.com/cpfair/quran-tajweed*; accessed September 2019.

5 Muṣḥaf Digital Sound

After the book of Labīb Al-Saʿīd (1969) on his efforts in making phonograph records of Quranic recitations, there has not been any documentation of the subsequent developments of Quranic sound records. Compiling such a chronology would require some extensive field search, looking through archives of the different media record companies and government organizations. Due to the absence of literature, it is not clear whether there has been any form of regulation through religious and governmental authorities in publishing quranic recitations in sound media, or whether it has been entirely community and commercial efforts. At present, the internet is the principal medium for storing retrieving all sound media. There are thousands of websites and platforms on the internet for accessing Quranic recitations; those are either dedicated websites or platforms for sharing media. Media Sharing platforms YouTube, SoundCloud, and the Internet Archive house massive collections of Quranic recitations and Quran related media: lessons, *tafsīr*, and recitations in other languages.

Certain repositories represent a more valuable resource for researchers or developers working on developing specialized software. These repositories usually utilize better organizational paradigms, extensive meta-data schemas, and house more massive, more diverse collections regarding reciters and recitation forms *qirāʾāt*. Following is a break-down of the most mention-worthy collections.

5.1 Data Reposteries of Quranic Recitations

5.1.1 IslamWeb and IslamWay[31]

IslamWeb is an online platform that specializes in providing educational materials on Islam and is curated by the ministry of the Qatari Ministry of Awqāf. IslamWay has similar form but does not provide any credits on who the curators are. Both websites have an extensive collection of Quranic recitations in almost 25 different *rwāyāt* and from around 400 reciters.

5.1.2 Verse by Verse Quran[32]

This project provides a highly extensive indexing schema of its recitations data repository. It provides audio files that are indexed with intervals in seconds of each *āya* per *sūra* and reciter. Furthermore, individual *āyat* are also available as separate files. This data

[31] *islamweb.net* and *islamway.net*

[32] *versebyversequran.com*; accessed Sepetmber 2019.

repository has been utilized in multiple applications that provide easy access of recitations to users by directly choosing a certain *āya* without needing to run the recording of the sūra from the beginning or manually seek the playing forward to reach the desired point. Moreover, such elaborate indexing makes it easier to apply automated analysis procedure in applying speech analytics or aligning the performance of different reciters and comparing them against one another. The project website does not provide any credits and seems to be a community effort.

5.1.3 Word by Word Quran[33]

This project is another valuable resource for developers, as it provides a deeper level of indexing; that is a separate file for each word per *āya* and *sūra*. The files are straightforward pronunciation of the words and not in a specific recitation style. This data repository has been of great benefit for software that aims at teaching Quran recitations to non-native Arabic recitations; as the users can individually play and repeat the words that might be unclear or hard for them. Additionally, this data has been of great help for developers working on building acoustic models for Arabic and Quranic recitation; which usually involves tokenizing an audio file into individual words. This collection spares this first step by providing the words already tokenized. The primary deficiency with this project that it only provides so far single recording per word; for better acoustic models a bigger, more diverse sets of records are needed.

5.2 Software and Methods for Automatic Processing of Recitations

5.2.1 Discriminative Training for Phonetic Recognition of the Holy Quran

Baig, Qazi, and Kadri (2015) apply a set of machine learning techniques to design a method of recognizing Quranic recitations. This method dissects recitation into separate phonemes which are then mapped into words. Whether the recognition is successful depends on matching the resulting text to Quranic digital text. The study was carried out by constructing acoustic models from ten different recorded recitations and one tester. Final results had accuracy rates up to 84.64%.

5.2.2 Building CMU Sphinx language model for the Holy Quran

El-Amrani et. al (2016) did an experiment with a speech recognition software library called CMU Sphinx and Quranic recitations. They carried out the experiment with recitations from 21 reciters of 4 short *swar*. The experiment was highly successful, although

[33] *github.com/quranwbw/audio-words-new*; accessed Sepetmber 2019.

generalizing it would be effortful, as it would involve generating phonetic transcriptions for the complete Quran; ideally in all recitation forms *qirāʾāt*.

5.2.3 Tarteel.io[34]

This project successfully achieves the goals that the previously mentioned methods aimed at reaching. Tarteel.io is a community effort, and it is unclear whether it has been dependent on any scholarly work or either of the studies so far mentioned. In contrast to the scholarly articles, no explanation of the methodology is provided, and only the software itself is presented. Using this application, it is possible to perform voice search in the Quran; i.e., reciting an *āya* and getting the text and the exact position of its occurrence in the Quran. The project aims at providing advanced analytics of Quranic recitations; however, its curators have not specified any precise details on the desired analytics. The advancement of the project is dependant on crowdsourced data collection of recited Quran of professional and non-professional reciters.

[34] Infromation from *tarteel.io/about*; accessed September 2019.

6 Conclusion

The curation of the Quran in the digital age seems to be a continuation of how this has been carried out over the past centuries: not only by scholars and authorities but by the engagement the Muslim community as well. The details of how the Quran came to be digitized and widely available on the internet and through diverse platforms and devices is unclear; there are no traces of this widespread having been a result of large-scale government-funded project and was most probably carried out to the largest extense by communities of devout Muslims. The same applies to advanced software that involves elaborate processing of Quranic script and recitations: communities have been more successful than scholars in producing functional software with satisfactory results. Nonetheless, scholarly work has been active in producing theories and methods in producing better Quranic Software; mainly e-learning software. There has been an emerging field of research labeled as Digital Quran Computing[35]. Many of the methods and theories available in the literature remain without application.

The advancements in speech analytics seem promising in terms of the potential of giving better insights on how performances of Quranic recitations vary. The main obstacle in achieving this goal is the lack of collaboration between scholars of Quranic Studies and researchers developing computational methods. The current approach in digitizing Quranic speech emphasizes constructing acoustic models that aggregate the similarities among recitations so that these models can be matched to all recitations. There have not been any efforts in producing such models of specific features of recitations; like tajweed phenomena that would be of more interest for Scholars of Quranic studies and inspect and differentiate.

[35] Zakariah et. al. (2017).

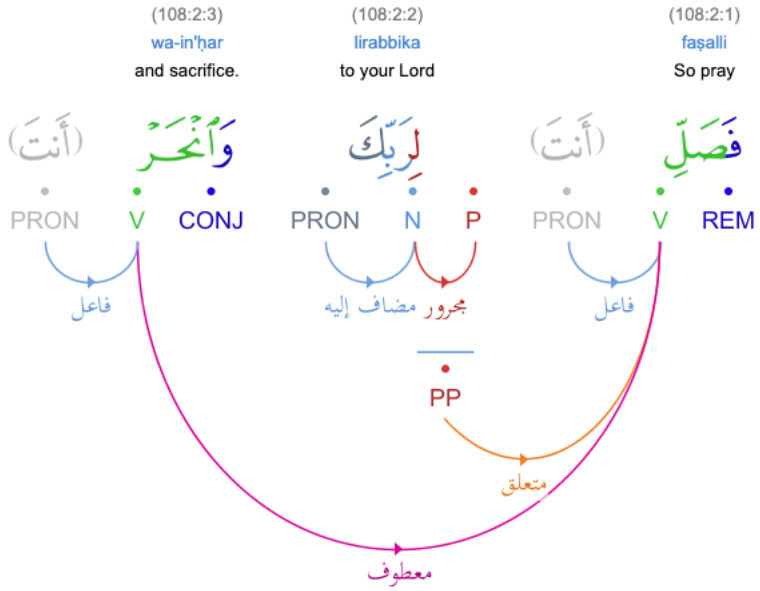
7 Appendix

Example from Oujda Quranic Corpus:

```
<Chapter Name="سُورَةُ الْكَوْثَرِ" Number="108"    locale="مكية">
    <Word Verse_number="0"  is_basmalla="1" Vowledform="بِسْمِ"    Stem="سم"   StemPattern="ٱفع"  PoS="اسم جامد"  Lemma="اسْمٌ"    LemmaPattern="ٱفع" Root="سمو"/>
    <Word Verse_number="0"  is_basmalla="1" Vowledform="ٱللَّهِ"   Stem="الله" StemPattern="#" PoS="اسم الجلالة" Lemma="الله"   LemmaPattern="#"   Root="#"/>
    <Word Verse_number="0"  is_basmalla="1" Vowledform="ٱلرَّحْمَٰنِ"  Stem="رحمن" StemPattern="فَعْلَان" PoS="اسم جامد" Lemma="رَحْمَان" LemmaPattern="فَعْلَان" Root="رحم"/>
    <Word Verse_number="0"  is_basmalla="1" Vowledform="ٱلرَّحِيمِ"  Stem="رحيم" StemPattern="فعيل" PoS="صفة/اسم الفاعل" Lemma="رَحِيم" LemmaPattern="فعيل" Root="رحم"/>
    <Word Verse_number="1"  is_basmalla="0" Vowledform="إِنَّا"    Stem="انا"  StemPattern="#" PoS="حرف تنبيه" Lemma="إِنَّا"  LemmaPattern="#"   Root="#"/>
    <Word Verse_number="1"  is_basmalla="0" Vowledform="أَعْطَيْنَٰكَ" Stem="عطين" StemPattern="أفعلنا" PoS="فعل ماض مبني للمعلوم" Lemma="أَعْطَى" LemmaPattern="أفعل" Root="عطو"/>
    <Word Verse_number="2"  is_basmalla="0" Vowledform="ٱلْكَوْثَرَ"  Stem="كوثر" StemPattern="#" PoS="اسم علم" Lemma="كَوْثَر" LemmaPattern="#"   Root="#"/>
    <Word Verse_number="2"  is_basmalla="0" Vowledform="فَصَلِّ"   Stem="فصل" StemPattern="فَعِّل" PoS="فعل أمر" Lemma="صَلَّى" LemmaPattern="فَعَّل" Root="صلو"/>
    <Word Verse_number="2"  is_basmalla="0" Vowledform="رَبِّكَ"   Stem="ربك" StemPattern="فعل" PoS="اسم جامد" Lemma="رَبّ"  LemmaPattern="فعل" Root="ربب"/>
    <Word Verse_number="2"  is_basmalla="0" Vowledform="وَٱنْحَرْ"  Stem="انحر" StemPattern="افعل" PoS="فعل أمر" Lemma="نَحَرَ" LemmaPattern="فعل" Root="نحر"/>
    <Word Verse_number="3"  is_basmalla="0" Vowledform="إِنَّ"    Stem="ان"   StemPattern="#" PoS="حرف توكيد تشبيح" Lemma="إِنَّ"  LemmaPattern="#"   Root="#"/>
    <Word Verse_number="3"  is_basmalla="0" Vowledform="شَانِئَكَ"  Stem="شنىء" StemPattern="فَاعِل" PoS="اسم جامد" Lemma="شَانِئ" LemmaPattern="فَاعِل" Root="شنء"/>
    <Word Verse_number="3"  is_basmalla="0" Vowledform="هُوَ"    Stem="هو"   StemPattern="#" PoS="ضمير الغائب – للمفرد المذكر" Lemma="هو"  LemmaPattern="#"   Root="#"/>
    <Word Verse_number="3"  is_basmalla="0" Vowledform="ٱلْأَبْتَرُ"  Stem="ابتر" StemPattern="أفعل" PoS="اسم جامد" Lemma="أَبْتَر" LemmaPattern="أفعل" Root="بتر"/>
</Chapter>
```

Example Annotation from Leeds Quran Corpus

Chapter (108) sūrat l-kawthar (A River in Paradise)



Translation	Arabic word	Syntax and morphology
(108:2:1) faṣalli So pray	فَصَلِّ V REM	**REM** – prefixed resumption particle **V** – 2nd person masculine singular (form II) imperative verb الفاء استئنافية فعل أمر
(108:2:2) lirabbika to your Lord	لِرَبِّكَ PRON N P	**P** – prefixed preposition *lām* **N** – genitive masculine noun **PRON** – 2nd person masculine singular possessive pronoun جار ومجرور والكاف ضمير متصل في محل جر بالاضافة
(108:2:3) wa-in'ḥar and sacrifice.	٢ وَٱنْحَرْ V CONJ	**CONJ** – prefixed conjunction *wa* (and) **V** – 2nd person masculine singular imperative verb الواو عاطفة فعل أمر

8 Bibliography

Alsmadi, Izzat, and Zarour Mohammad, Online integrity and authentication checking for Quran electronic versions, Applied Computing and Informatics, Vol. 13, 2017,

El Amrani, Mohamed Yassine; Rahman, M.M. Hafizur; Wahiddin, Mohamed Ridza; and Shah, Asadullah, Building CMU Sphinx language model for the Holy Quran using simplified Arabic phonemes, Egyptian Informatics Journal, Vol. 17, 2016.

Baig, Mirza Muhammad Ali; Qazi, Saad Ahmed; and Kadri, Muhammad Bilal. Discriminative Training for Phonetic Recognition of the Holy Quran. Arabian Journal for Science and Engineering, Vol. 40, 2015.

Bentrcia, Rahima; Zidat, Samir; and Marir Farhi, Extracting semantic relations from the Quranic Arabic based on Arabic conjunctive patterns, Journal of King Saud University - Computer and Information Sciences, Volume 30, 2018.

Bountouri, Lina. Archives in the Digital Age: Standards, Policies and Tools. Chandos Publishing, 2017.

Dukes, Kais. Statistical Parsing by Machine Learning from a Classical Arabic Treebank. Doctoral Thesis, The University of Leeds School of Computing, 2013

Guellil, Imane; Saâdane, Houda; Azouaou, Faical; Gueni, Billel; and Nouvel, Damien, Arabic natural language processing: An overview, Journal of King Saud University - Computer and Information Sciences, 2019,

Hirschkind, Charles. "Media and the Quran." In Encyclopedia of the Quran, Vol. 3. Leiden-Boston: Brill 2003.

Jurafsky, Dan, and Martin, James H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Prentice Hall, 2009.

Al-Saᶜīd, Labīb. Al-ǧamᶜ al-ṣawtī al-awal lil-qurʾān al-karīm. Dār al-kitāb al-ᶜarabī lil-ṭibāᶜa wan-našr, Cairo 1969.

Zeroual, Imad, and Abdelhak Lakhouaja. "A New Qurʾānic Corpus Rich in Morphosyntactical Information." International Journal of Speech Technology 19, no. 2 June 2016.

Zakariah, Mohammed; Khan, Muhammad Khurram; Tayan, Omar; and Salah, Khaled.

Digital Quran Computing: Review, Classification, and Trend Analysis. Arabian Journal for Science and Engineering, Vol. 42, 2017.