Journal of Semitic Studies LXIII/2 Autumn 2018 doi: 10.1093/jss/fgy005

AUTOMATIC EXTRACTION OF QURANIC LEXIS REPRESENTING TWO DIFFERENT NOTIONS OF LINGUISTIC SALIENCE: KEYNESS AND PROSODIC PROMINENCE

CLAIRE BRIERLEY[1] MAJDI SAWALHA

UNIVERSITY OF LEEDS UNIVERSITY OF JORDAN

TAJUL ISLAM, JAMES DICKINS and ERIC ATWELL

UNIVERSITY OF LEEDS

Abstract

This paper presents two sets of lexical items automatically extracted from the Arabic Qur'ān, and denoting two different notions of linguistic salience: *keyness* and prosodic *prominence*. Our novel hypothesis investigates a possible correlation between them. Our novel findings discover distributionally significant keywords that also occur strategically in phrase-final position so as to maximise their prominence, and thus meaningfulness, for reader, reciter, and aural recipient. Our methodology first computes Quranic keywords via the Corpus Linguistics technique of Keyword Extraction, and maps them to major Quranic themes in Islamic scholarship. Next, we implement a bespoke algorithm for rule-based capture of words annotated with *madd* or prolongation, a specific type of prosodic highlighting in Quranic recitation rules or *tajwīd*. We find it especially interesting that the concept of final syllable lengthening (*madd* before pause) is encoded in *tajwīd* and effectively demarcates phrase boundaries in the Qur'ān. We concentrate on nominal keywords (*i.e.* nouns and adjectives) since these are more likely to be aligned with phrase edges and to bear the hallmarks of pre-boundary lengthening. This correlation between keyness and prominence occurs 43.29% of the time in our data, since 526 keywords appear in our extracted subset of nominal types tagged with *madd* before pause: ((526/1215)*100). Finally, we identify *which* Quranic keywords are most likely to be annotated with enhanced prolongation in the final syllable before pause, using an easy-to-interpret, single value metric: the *Laplace Point Estimate*. Keywords that emerge as semantically weighted in terms of both distributional and prosodic significance are most likely to reflect the Quranic themes of *God*, *Nature*, and *Eschatology*.

[1] Principal author, lead researcher and corresponding author.

1. Introduction

In this paper, we describe and implement two approaches for automatic extraction of meaningful terms in the Arabic Qur'ān. These approaches presuppose two different notions of linguistic salience – that associated with statistical significance or *keyness* (Scott 1997) versus that associated with prosodic prominence — and yield two corresponding item sets. We compute the first set via the Corpus Linguistics technique of Keyword Extraction (KWE) which identifies words of unusual frequency in a test corpus (in this case the *Qur'ān*) versus a suitable reference corpus. We extract the second set via rule-based capture and frequency profiling of words marked as prosodically prominent according to the rules of traditional Quranic recitation or *tajwīd*. Finally, we focus on overlap between these sets, and propose a metric for measuring this overlap. This is original research since a potential correlation between distributionally significant and prosodically prominent words is hitherto unexplored both in the Qur'ān and in Corpus Linguistics more generally. We explore this hypothesis via corpus-based (quantitative) and qualitative approaches. Also, by juxtaposing these two different notions of salience (i.e. keyness and prominence), we hope to gain further insight into the prosody-semantics interface.

We postulate that keywords extracted via computation over the Qur'ān may already be encoded as meaningful in Quranic *tajwīd*. *Tajwīd* has evolved as a discipline within Islamic Studies, and an annotation system for the Qur'ān to facilitate clear articulation and phrasing during a recitation of the scripture. Central to this paper is the concept of *prolongation* or *madd* manifest in *tajwīd* mark-up of nuclei in lengthened syllables, since enhanced duration has long been accepted as an acoustic cue to prominence (Mishra et al. 2012; Klatt 1976). Another important aspect is prescriptive mark-up of pauses or *waqf* within and between verses, especially since pre-boundary lengthening is implicitly identified in a subset of *madd*. Our *madd* extraction algorithm operates over plain text and phonemic transcriptions to identify and count instances of all the different types of *madd* in the Qur'ān, thus enabling comparison of our subset with extracted keywords. Language processing is accomplished in Python and Python-based Arabic language processing tools.

This interdisciplinary investigation necessitates discussion of English as well as Arabic linguistic theory since it is through the former that key concepts linking prosody and information status have gained recognition and acceptance. These are covered in depth in Section 2

of this paper. Theoretical discussions continue in Sections 3 and 4 which cover prolongation in Quranic *tajwīd*, and the likelihood ratio test for statistical significance and corpus comparison respectively. To maintain continuity for the reader, we then report on KWE in two consecutive sections, dealing with corpora used in our experiments and corpus pre-processing in Section 5, and detailed analysis and interpretation of extracted keywords in Section 6. In Section 7, we present our algorithm for codification of the different types of *madd*, and extraction and frequency profiling of words in two subsets of pre-boundary words. The resulting lexical items are then discussed in relation to our keyword lists (Section 8), and conclusions drawn (Section 9).

2. Prosody and Information Status

There is an extensive body of literature on prosody *per se* and on the information-bearing aspects of perceptually prominent words. Prosody has been studied from a variety of theoretical perspectives and experimental approaches, including: phonetics; phonology; syntax; and discourse. In this section, we draw on this diversity to review key concepts in the realization and interpretation of prosodic prominence and phrasing in relation to information status in both English and Arabic as a basis for interpreting *tajwīd* annotation of *madd* (prolongation) on syllable nuclei in Quranic Arabic.

2.1 Realisation of Prosodic Prominence: Evidence from English Phonetics

Prosodic prominence in speech is measured in terms of acoustic and articulatory phenomena such as energy, duration, and fundamental frequency (Mishra et al. 2012). In a landmark study in the field of speech communication for English, Klatt (1976) associates prominence with duration at various segmental levels (phonemes, syllables, and words), and durational patterns with meaningfulness. Variability in segmental timing is partly due to the intrinsic properties of individual phonemes, but is also influenced by stress, emphasis, and positional factors, namely: the location of a phoneme within a word, and the location of a word within an utterance. Increased duration due to stress, emphasis and/or position largely affects vowel segments and some postvocalic consonant segments (*i.e.* sonorants and fricatives).

Stress, emphasis, and position are interrelated. The location of word stress in English is said to be variable; but stressed syllables are typically perceived as being louder and longer than others, having a full vowel not a reduced one. Stressed syllables may also acquire extra emphasis under certain constraints; this is realized via fluctuations in pitch, hence the term pitch accent. In another landmark, phonetics-based study by Fry (1958), the most significant correlate of perceived prominence was found to be higher frequency (F_0). The relationship between stress and pitch accents is widely accepted since the latter are assigned to stressed syllables within a word. The *inter*relationship between stress, pitch accents, and location then operates at phrasal level. Gordon (2014) affirms a dominant, cross-linguistic pattern of promoting the primary stressed syllable of the rightmost content word within an intonational phrase (IP) to the status of pitch accent in default prosodic conditions; and aligning pitch accents with primary stresses is also habitual in Arabic (de Jong and Zawaydeh 1999). Wichmann *et al* (2009) define an IP or 'chunk' as a sequence of (transcribed) speech that contains at least one pitch-accented word. Where there is more than one tonal prominence within the IP, the rightmost content word attracts the most prominent pitch accent (Ladd 1996 in Gordon 2014).

2.1.1 Pre-Boundary Lengthening

Final syllable lengthening is a prosodic device in many languages, including English (Klatt 1976) and Arabic (De Jong and Zawaydeh 1999), whereby a segment/phoneme in word-final position is demonstrably longer than that same segment/phoneme in word-medial position. Wightman et al. (1992) further specify that pre-boundary lengthening occurs in the rhyme of the final syllable, namely: the vowel and subsequent consonant(s), but not the consonantal onset. They also find that this lengthening effect is both prevalent and highly informative since it is used cumulatively by speakers to distinguish different boundary types: segments in final position of larger prosodic domains are characteristically longer than those in final position of smaller domains (ibid.; Gordon and Munro 2007).

2.2 Interpretation of Prosodic Prominence in English Phonology and Discourse

Klatt (1976) associates prosodic prominence (including durational and positional features) with meaning. However, there is still controversy in the literature as to *how* prosody and phrasing relate to infor-

mativeness (Calhoun 2010a; 2010b). The main contention is whether prominence (e.g. the distribution of pitch accents) reflects prosodic or discourse structure (Hellmuth 2007).

2.2.1 The Prosodic Hierarchy

In the predominant Autosegmental-Metrical (AM) model of intonational phonology (Nespor and Vogel 1986; Selkirk 1986), the so-called prosodic hierarchy defines a series of phonological units below the level of an utterance, namely (in descending order): intonational phrase (IP); phonological phrase (PhP) or (φ); prosodic word (ω); foot (F); syllable (σ); mora (μ); (and possibly) segment. Phonological phrases may be further decomposed into: (i) an intermediate or major phonological phrase; and (ii) an accentual or minor phonological phrase (Selkirk 2005). Stress and pitch then relate to meaning at different levels of this hierarchy (Frota et al. 2012), where, in general terms, stress operates at word-level, and pitch at the level of phrasing (Gordon 2014).

2.2.2 Prosody and Phrasing

Intonational and phonological phrases are important concepts for understanding the syntax-phonology interface and can be realized and detected phonetically via 'prominent' tonal and durational cues such as: phrase (edge) tones; boundary tones; pitch reset; pre-boundary (i.e. final syllable) lengthening; and pause. However, for English, complexity is introduced by the co-occurrence of what are variously identified as phonetic, metrical, syntactic, and pragmatic entities at/near rightmost phrase edges. English is constrained by its metrical structure to mark the head of every phonological phrase with a pitch accent (Hellmuth 2014; Calhoun 2010b; Wichmann et al. 2009), where the nuclear or most prominent pitch accent is habitually reserved for the final phrase in the utterance. Furthermore, pitch accented words frequently coincide with phrase and boundary tones since they tend to be phrase-terminal. This is turn overlaps with the right-branching bias in syntactic structure, where tonal prominences frequently align with syntactic heads, and where syntax — conceptualized as content versus function words, or part-of-speech tags, or treebank-style parse features — has proven to be a reliable predictor of prosodic-syntactic phrase boundaries in automated phrase break prediction (Liberman and Church 1992; Taylor and Black 1998; Read and Cox 2007).

2.2.3 The Discourse Functions of Prosody

Finally, prosodic prominence and phrasing are thought to signal information structure, where key concepts are: (i) focus (F-marking); (ii) givenness; and (iii) and the theme-rheme (or topic-comment) construct (Gundel 1988; Halliday 1967). In the analysis of discourse, mentions of an entity are either 'given' or 'new': they either refer back to, or update the existing discourse model. New information is F-marked, and pitch accents are thought to be functionally motivated in highlighting salience (newness). Prosodic phrasing then relates to the organization of information within an utterance. The distribution of F-marked items is considered in terms of theme-rheme, namely: the previously established (given) topic or theme, and new commentary on that topic (the rheme).

There is broad consensus that prosodic prominence and phrasing are crucial to conveying meaning. Prosody is used to partition the speech stream into chunks of constituents that can be interpreted together (phrasing), and highlight new and important content within each chunk (prominence). One characteristic of English is that it likes to chunk and highlight at the same time (Brierley and Atwell 2007); due to right-branching bias, prominent words tend to complete a phrase group. The traditional view is that pitch accents are functionally motivated and mark focus: this is termed the 'accent-first' or 'pitch-first' approach (Hellmuth 2007; Calhoun 2010a). However, prominence is also constrained by metrical structure; accents emerge at strong nodes within a hierarchical, binary-branching system of weak versus strong syllables. This is the 'stress-first' approach (Hellmuth 2007): pitch accents are, first and foremost, phonologically-determined events. For a language like English, the analysis of form (stress-first) coincides with the analysis of function (pitch-first) since pitch accents mark phrase level prominence and hence interact with focus.

2.3 Realisation and Interpretation of Prosodic Prominence in Arabic

Arabic is similar to English in that it is stress-timed (Nespor et al. 2011; Jun 2005). It has pitch accents aligned to primary stresses (de Jong and Zawaydeh 1999) and to constituents within the prosodic hierarchy (Hellmuth 2007). The prosodic typology of Arabic is in fact very similar to that of English (Jun 2005). In both languages, prominence is marked by stress at the lexical level, and demarcation of the heads of prosodic units at the post-lexical level (ibid.). Furthermore, rhythmic units in both languages are conceived of as the

metrical foot (a grouping of strong and weak syllables) at level of lexis, and the intermediate/phonological phrase (PhP/φ) and intonational (IP) phrase at the post-lexical level (*ibid*).

Lexical stress is much more predictable in Standard and Classical Arabic than in English, however, being a function of syllable weight and position (Ryding 2014; Watson 2011). In 'full-form' (i.e. fully-inflected) pronunciation, primary stress in Arabic falls on: (i) the first syllable of disyllabic words; (ii) the penultimate syllable if it is heavy or (failing that) the antepenultimate syllable in polysyllabic words (Ryding 2014). Stress increases vowel durations in Arabic as in English, but quantity lengthening/contrast in Arabic is primarily phonemic: it is a mechanism for differentiating the long and short vowels (de Jong and Zawaydeh 2002). As we shall see in Section 3, prolongation in Quranic *tajwīd* only affects the long vowels (ا و ي). Furthermore, since the rules of *tajwīd* recitation are there to ensure clarity of perception as well as production, the device of prolongation as a form of 'hyperarticulation' (de Jong 1995) serves to maximise vocalic contrasts. This in turn has been noted as an overarching feature of stress-timed versus syllable-timed languages: the former exhibit more variation (contrast) between two consecutive vowel durations (Jun 2005).

In a landmark study of durational, spectral and F_0 correlates of stress and word-final juncture in Jordanian Arabic, de Jong and Zawaydeh (1999) found that the realisation of Arabic word-level prosody is very similar to that of English. Most notably, they observed extensive pre-boundary lengthening effects as distinct from stress. They also observed tonicity (pitch accents) associated with stressed syllables, and juncture-marking phrasal pitch specifications (tonality). Their conclusions reassert the cognitive function of prosodic prominence and phrasing as a language universal. Furthermore, in a later study (de Jong and Zawaydeh 2002), they engage with the discourse notion of focus, and its realisation at syllable, word, and phrase level; and they again differentiate stress from positional lengthening: while both produce longer durations, only stress produces more extreme formant values.

3. Quranic *Tajwīd* and *Madd*

The Qur'ān as a scriptural document has undergone some developmental changes which ultimately lead us to the standard copy (*muṣḥaf*) that we have today. Five distinct phases can be identified with the progression of the Quranic text, all of which are significant for

Quranic *tajwīd*. The first phase begins with the 'master copy' or Uthman Codex (*al-rasm al-ʿuthmānī*) which was initiated to unite all Muslims upon one standardized 'book'; this may be termed 'codexing' (*tarsīm*). The next phase ushers in the use of diacritical points; this may be termed 'consonantisation' (*iʾjām*), since the application of such provides clarity to pronunciation of letters. The third phase involves the introduction of short vowels (*ḥarakāt*); this can be termed 'vowelisation' (*tashkīl*), which offers clarity to the pronunciation of words. The fourth phase introduces orthographic signs, pauses (*waqf*) to ensure correct meaning, and numerals to separate the verses. This may be termed 'punctuation' (*tarqīm*) and is most significant for the rules of *madd*. The fifth and final phase is the crystallization of all the previous phases into the set of *tajwīd* recitation rules. The attempts of *tajwīd* experts to codify and objectively reproduce the original, 'revelatory' pronunciations demonstrate that standardization of the Quranic text did not end with Uthman's initiative.

The *tajwīd* recitation rules seek to encapsulate and facilitate reproduction of canonical (and indigenous) pronunciations of Arabic phonemes passed down through the oral tradition (Shah 2003). At the same time, they represent ancient, insightful, linguistic analyses by a handful of trained Quranic reciters or *qurrāʾ*, plus early Arab grammarians such as Al-Khalīl (d. [around] 170/786), Sibawayh (d. [around] 180/796), and (later) Al-Khāqānī (d. [around] 325/936) which are pertinent to modern, Western linguistic theory (ibid.; Heselwood and Hassan 2011; Brierley et al. 2016). Giving each letter its 'rights' and its 'dues' during Quranic recitation means pronouncing each Arabic phoneme perfectly in terms of place and manner of articulation (its rights), plus coarticulatory context (its dues), so as to distinguish it from any other phoneme in the language and thus preclude error or distortion in transmission. The subset of *tajwīd* rules governing vowel durations is known as *madd* or prolongation.[2]

3.1 Overview of the Different Types of Madd

Quranic *tajwīd* defines three overarching branches of *madd*[3] or prolongation of syllable nuclei and these are: (i) natural *madd*; (ii) *madd*

[2] Tradition has it that Qatāda asked the companion Anas Ibn Mālik: 'How was the Prophet's recitation?' — to which Anas replied: 'He used to elongate [his voice]', *kāna yamuddu maddan* (Al-Masīrī 2002: 138).

[3] A diagrammatic representation of the different branches of *madd* can be found on the *Read with Tajweed* website: http://www.readwithtajweed.com/images/Charts/MaddChart.pdf

caused/followed by *hamza*; (iii) *madd* caused/followed by *sukūn*. Natural *madd* letters are the long vowels *'alif*, *wāw*, and *yā'*: ا و ي. These are twice as long as short vowels and equal to two 'counts' whenever they are *not* preceded by *hamza* and/or *not* followed by *hamza* or a silent letter. In most *tajwīd* manuals, a count is said to be roughly equal to the time it takes to unfold one finger from a lightly clenched position (and two counts to the time it takes to say *'alif*). In terms of phonology, counts are synonymous with morae, units denoting syllable weight. Hence Arabic diphthongs *aw*, ـَوْ and *ay*, ـَيْ are also prolonged for two counts during recitation, where the same contextual constraints apply for *hamza* and *sukūn*.

A related case is normal prolongation (2 counts) of the short vowel in the suffixed masculine singular object pronoun *-u* ـهِ in continuous recitation, as long as the next word does not begin with a silent letter or with *hamza*. In most cases, the short vowel in this suffix is *ḍamma* as in Q.43.4: وَإِنَّهُ فِي أُمِّ الْكِتَابِ (*wa innahū fī ummi l-kitābi, And indeed it is, in the Mother of the Book*); in other cases it is kasra as in Q.11.8: كَانُوا بِهِ يَسْتَهْزِئُونَ (*kānū bihī yastahzi'ūna, what they used to ridicule*). Another related case is realisation of *tanwīn al-fatḥa* before pause, where it is pronounced as a normal *'alif*, as in the phrase عَلِيْمَا قَدِيْرَا (*ʿalīman qadīrā*, *He is ever Knowing and Competent*).

3.1.1 Abnormal Prolongation Associated with Hamza

In Quranic recitation, the duration of long vowels and diphthongs is prolonged for four, five or six morae depending on context; this is termed abnormal prolongation. Four scenarios arise in the vicinity of *hamza*, the glottal stop.

Prolongation of the long vowel or diphthong as syllable nucleus for at least four counts is obligatory if it is immediately followed by *hamza* in the same word. In some cases, when *hamza* is the last letter of the word and the reciter pauses after it, it is extended to six counts. For example, both words are prolonged in the phrase هَنِيئًا مَّرِيئًا *hanī'an mmarī'ā*, *in satisfaction and ease*, followed by pause.

While the reciters (*qurrā'*) are in agreement regarding the effect of *hamza* on prolongation, they differ on the extent. Warsh (d. 197/813) and Ḥamza (d. 156/773) both argue that it should last for six short vowels, ʿĀṣim contends (d. 127/745) that it should last for five short vowels, and finally al-Kisā'ī (d. 189/805) maintains that it should last for four short vowels (al-Ḍabbāʿ 1997: 97). These differences are typical and probably reflect different, authorized recitation styles and speeds. However, the important point to make is that such 'abnormal' durations can be interpreted as a more or less conscious device

to mitigate anticipatory co-articulation triggered by the glottal stop, which might otherwise result in a tendency to foreshorten the vowel sound, leading to impure realization. Furthermore, hyperarticulation of the long vowel in the final syllable before the Arabic phonemes /m/ and /n/ may serve to enhance *ghunnah* or nasal humming (another *tajwīd* category) in phrase/verse terminal words like مستقيم *mustaqīm* (Q.1.5) and العالمين *al-ʿālamīn* (Q.1.2).

In continuous recitation, it is permissible to prolong a long vowel or diphthong for four or five counts at the end of a word if the next word begins with *hamza*. An instance of this type is Q.108.1: إِنَّا أَعْطَيْنَاكَ الْكَوْثَرَ *innā aʿṭaynāka al-kawṯara,* 'Indeed, We have granted you, [O Muhammad], al-Kawthar'.[4]

Section 3.1 outlines how the short vowel in the suffixed masculine singular object pronoun gains an extra count in continuous recitation as long as that word does not begin with *hamza* or a silent letter. If the subsequent word does happen to begin with *hamza*, the count on this suffix is (preferably) extended to four or five morae to distinguish it from naturally occurring *madd* (*Read with Tajweed* 2016). This again hints at the linguistic motivation for 'abnormal' prolongation of long vowels, namely: preservation of vocalic contrasts during co-articulation. An example would be Q.2.112: فَلَهُ أَجْرُهُ عِندَ رَبِّهِ (*falahū ağruhū inda rabbihi*, *will have his reward with his Lord*). Once again, there are differing views among reciters as to duration for this type of madd (al-Ḍabbāʿ 1997: 98–9).

Exchange of order prolongation occurs when the *madd* letter is immediately preceded by *hamza* in the same word. It is really a case of normal prolongation since the nucleus is not extended beyond two counts. However, it is categorised here because the *madd* letter is a substitute, introduced for ease of pronunciation, for what was originally a *hamza* (*Read with Tajweed* 2016). For example, *hamza* precedes *wāw* in the word أُوذِيْنَا (*we have been harmed*). This is a natural *madd* and does not exceed two short vowels (al-Masīrī 2002: 153). It is read in continuance and pause (Ghawthānī 1996: 40).

3.1.2 Abnormal Prolongation Associated with Sukūn

Madd letters are prolonged for the duration of six morae if they precede *sukūn*, namely: a silent letter within the same word or a pause. This rule also affects realization of disconnected letters heading many chapters in the Qur'ān, for example: قٓ وَالْقُرْآنِ الْمَجِيدِ (*qāf wal-qur'ani al-mağīdi*, 'Qaf. By the honoured Qur'ān' (Q.50.1).

[4] *al-Kawthar* is the name of a river in Paradise.

There is a single word-internal case where a long vowel precedes an original silent letter which is not doubled; this occurs in the word آلْآنَ *āl āna*, *now*, which is repeated twice in the Qur'ān. However, there are frequent occurrences of *madd* letters followed by an original silent and doubled letter as in: الضَّآلِّينَ *al-ḍāllīna*, 'those who are astray' (Q.1.7). This is further discussed in Section 3.2.1.

Finally, when a long vowel occurs in the syllable before pause, where the pause in question is marked in *tajwīd* annotation, and is not a disfluency, then it is prolonged for up to six morae. This is termed *madd al-āriḍ* (*the accidental stop*) and is of most interest for this paper. A diagrammatic representation[5] of the different branches of *madd* can be found on the *Read with Tajweed* website.

3.2 Prosodic Prominence Realised as Madd *before Pause*

The rules of *tajwīd* have evolved to eliminate any misunderstandings of the divine message resulting from incorrect or unclear pronunciation. As stated, they are derived from ancient practice: that of professional reciters (*qurrā'*) from as far back as the seventh century (Bohas et al. 1990; Shah 2003). From our contextual summary of *tajwīd* rules governing *madd*, we find that the motivation for this set is predominantly phonemic: to enhance the duration of syllable nuclei (a form of hyperarticulation) so as to maximize the quantity contrast differentiating long and short vowels (§2.3). However, we also find that the category of *madd* before pause (i.e. *madd al-āriḍ*, *the accidental stop*) does identify semantically salient words when that pause is explicitly marked in *waqf* annotation and is not simply a disfluency.

Madd before pause brings together key concepts discussed in Section 2, namely: (i) word-internal position affecting vowel duration, and increased duration interpreted as prominent and meaningful; (ii) pre-boundary lengthening signalling phrase edges and hence demarcating syntactic and information structure (i.e. sequences of words that should be interpreted together); (iii) right-branching bias in metrical as well as syntactic structure releasing prominent words at the end of a tone unit or phrase group; (iv) the likelihood that phrase-terminal syntactic heads will exhibit tonicity (pitch accents) as well as tonality (phrase tones) and pre-boundary lengthening, and thus constitute salient items: words that are 'in focus' or 'new' in discourse terminology. It is noteworthy that such phenomena have been observed and codified in Quranic *tajwīd*.

[5] http://www.readwithtajweed.com/images/Charts/MaddChart.pdf

3.2.1 Stress versus Prolongation: an Illustrative Example

We have seen that increased duration (prolongation) in the syllable before pause is specifically labelled in a subset of *madd* in Quranic *tajwīd*, linking this annotation subset with informativeness, and enhancing the information status of words bearing such annotations. Furthermore, although word stress and pre-boundary lengthening show a tendency to interact, they are distinct phenomena in Arabic as in English, and are separately coded in the *madd* schema. A clear instance of this can be found in Q.1.7: الضَّآلِّينَ (*the astray ones*). The word الضَّآلِّينَ is strategically positioned: not only is it phrase and verse terminal, it is also the last word in the opening chapter and semantically charged. We find that it is significantly prolonged during recitation since it is annotated with two types of *madd*. Word stress is located on the first syllable which is also a case of *madd* before *sukūn* associated with *šadda* (up to six counts), where the first doubled letter is effectively silent. The final syllable is then a case of *madd* before a different notion of silence, namely: pause (up to six counts).

3.3 Multimodal Demarcation of Madd

Distinctions such as lexical stress versus pre-boundary lengthening are visualised in *tajwīd* editions of the Qur'ān such as *Dar-Al-Maarifah* (2008); vowels may be colour-coded to identify different gradations or durations of *madd*. In Figure 1 the traditional colour codes (cf. Muhammad 2012: 59), are linked to the different types of *madd*.

Figure 1: Colour-coded categorisation of *madd* in a *tajwīd* edition of the Qur'ān

Dark Red	Necessary prolongation (up to 6 counts)
Blood Red	Obligatory prolongation (up to 5 counts)
Bright Orange	**Pre-boundary/final syllable lengthening** (up to 6 counts)
Cumin	Normal prolongation of long vowels (2 counts)

Pre-boundary words can accrue prosodic prominence via the licence to double or treble normal durations for the long vowels *'alif*, *wāw* and *yā'*. In Figure 2, we present three more examples of this phenomenon. We have mapped the original Arabic text to automatically generated, full-form[6] IPA phonemic transcriptions which have then been edited manually to imitate colour-coding on these same verses in

[6] Full-form as opposed to pause-form pronunciation as defined in Ryding (2014: 36–7).

Quranic *tajwīd* (cf. Brierley et al. forthcoming). Instances of *madd* before pause appear in bold and underline; one instance of *madd* before hamza appears in bold; natural *madd* sites are underlined; and letters/sounds subject to coarticulation appear in light grey font in Figure 2.

Figure 2: Visualising different categories of *madd* within full-form IPA transcriptions of Quranic verses in the Yusuf Ali English translation

Location	Original Arabic; IPA Phonemic Transcription; English Translation
Q.1.4	مَالِكِ يَوْمِ الدِّينِ
	ma**ː**liki-jawmi-ʔaddi**ː**ni
	Master of the Day of Judgement.
Q.101.5	وَتَكُونُ الْجِبَالُ كَالْعِهْنِ الْمَنْفُوشِ
	wataku:nu-ʔalʤiba:lu-kalʕihni-ʔalmanfu**ː**ʃi
	And the mountains will be like carded wool.
Q.12.44	قَالُوا أَضْغَاثُ أَحْلَامٍ وَمَا نَحْنُ بِتَأْوِيلِ الْأَحْلَامِ بِعَالِمِينَ
	qa:l**u:**-ʔadˤɣa**ː**θu-ʔaħla**ː**min-wama:-naħnu-bitaʔwi:li-ʔalʔaħla**ː**mi-biʕa**ː**limi**ː**na
	They said: "A confused medley of dreams: and we are not skilled in the interpretation of dreams."

There are several points to notice here. First, colour coding of natural *madd* (where long vowels last twice as long as short vowels) only appears in certain contexts, namely: it is always marked on ˈ *dagger 'alif*, presumably because this glyph is associated with the Othmani script and may be less familiar to modern-day readers used to Modern Standard Arabic (MSA). The word مَالِك in Q.1.4 is rendered as مٰلِك in Othmani script; hence natural *mād* in the first syllable is colour-coded as a reminder (*Dar-Al-Maarifah* (2008). Naturally-occurring *madd* is not colour-coded in Q.101.5 however, presumably because readers will already be sensitive to this. The verse has been included to show the semantic weight (information status) of phrase-final and verse-terminal words in this powerful text. In the Quranic vision of entropy, mountains will be reduced to rubble: they will be like wool *fluffed up*: الْمَنْفُوش.

Verse 12:44 in the chapter entitled *Yusuf* is especially interesting. We see from the mark-up on /biʕa**ː**limi**ː**na/ that pre-boundary annotation of *madd* (bold and underline) occurs on what would be the penultimate syllable if the word retains its case ending. However,

since this word is verse-terminal, the case-ending is dropped, and pre-boundary lengthening coincides with primary stress on the final, super-heavy (CVVC) syllable. This is distinguished from normal prolongation on the *dagger 'alif* site in the mark-up (*i.e.* underline only).

Yusuf Ali's English translation of this verse (Q.12.44) uses punctuation to identify three separate phrases: '*They said: "A confused medley of dreams: and we are not skilled in the interpretation of dreams"*. It transpires that in the *tajwīd* version of this verse, the third word أَحْلَام is associated with an in-verse *waqf* boundary symbol ۖ. This denotes a permissible (though weak) stop. Thus, if the reciter chooses to pause here, the *tanwīn* kasra on أَحْلَامٍ would be dropped and the final syllable lengthened /ʔaħla̰ːm/.

4. Computing Statistically Significant Keywords

Keyword Extraction (KWE) is a statistical technique for corpus comparison used in Corpus Linguistics and is revelatory of significant lexical differences between two datasets in terms of *aboutness* and style (Scott 1997). It is exemplified in Text Analytics toolkits such as *Wordsmith Tools* (*ibid*) and *Wmatrix* (Rayson 2008). The methodology entails formal comparison, via significance testing, of observed word frequencies in a dataset of interest (*i.e.* the test set) with their expected frequencies in a suitable reference corpus, to verify the apparent overuse or underuse suggested by raw frequencies. Keywords in this context differ, therefore, from common parlance where the notion of *keyness* denotes words viewed *subjectively* as key; instead, they are words of unusual frequency (or infrequency) in a test set relative to a reference set, where unusualness depends on some predetermined confidence level of statistical significance. While results from KWE still require qualitative evaluation, the methodology mitigates against researcher bias since grammatical as well as content words may be retrieved. In our study, Keyword Extraction over the Qur'ān versus some 7 million words of Classical Arabic uncovers words/concepts that are quintessentially Quranic.

4.1 The Log Likelihood (LL) Metric

Both *Wordsmith Tools* and *Wmatrix* implement the likelihood ratio test for comparative frequency profiling and KWE over two corpora. This test is widely used in corpus studies since it does not assume a normal distribution (Rayson and Garside 2000) and is therefore more suitable for language data where an unlimited number of function

(grammatical) words abound in any one text/sample. The test compares the fit of two models, in this case the test and reference datasets, and computes the likelihood of the data under the former versus the latter. The resulting log likelihood (LL) statistic can also be used to compute a *p-value* denoting the probability of obtaining a test statistic at least as extreme as the one observed, assuming that there is no difference in the test and reference sets. In Corpus Linguistics, it is usual to set a very challenging LL statistic and p-value to determine whether there is a *significant* difference between any two datasets. *Wordsmith* and *Wmatrix* implement an LL cut-off of at least 6.63 for statistical significance; this corresponds to a p-value of 0.01 or 99% confidence that the difference (absolute value) between observed and expected values is not due to chance. This is the measure used in our experiments here.

Word frequency distributions are generated for both corpora (the test and reference sets), and the LL statistic is then calculated for each word in the respective frequency lists, following Rayson and Garside (2000) and *Wmatrix*.[7] First, expected frequencies are calculated from raw data drawn up in a contingency table (Table 1), where the values *a* and *b* represent individual word frequencies, and the values *c* and *d* represent the total number of words in the test and reference sets respectively.

Table 1: Contingency table adapted from *Wmatrix*: http://ucrel.lancs.ac.uk/llwizard.html

Observed Statistics in Data	Test Set	Reference Set	Total
Observed Frequency: Single Word)	a	b	a+b
Observed Frequency: All Other Words	c-a	d-b	(c+d)-(a-b)
Total	c	d	c+d

The expected frequency of any word in the test corpus is given by *E1 = c*(a+b) / (c+d)*, and the expected frequency of that same word in the reference corpus by *E2 = d*(a+b) / (c+d)*. Calculating LL can then be expressed as *2*((a*ln (a/E1)) + (b*ln (b/E2)))*, according to the formula:

$$2 \ln \lambda = 2 \sum_{i} O_i \ln \left(\frac{O_i}{E_i}\right)$$

Here *O* denotes *observed* frequency and *E* denotes *expected* frequency.

[7] http://ucrel.lancs.ac.uk/llwizard.html

5. Corpora used for Keyword Extraction

The test and reference sets used in this study are the Qur'ān versus the Literature section of the *King Saud University Corpus of Classical Arabic* (Alrabiah 2013) respectively. The Qur'ān corpus used is the *Boundary Annotated Qur'ān* (Sawalha et al. 2014; Brierley et al. 2012), henceforward referred to as BAQ. This dataset is purpose-built for machine learning, and features the whole text of the Qur'ān rendered in traditional Othmani script and Modern Standard Arabic (MSA), with Arabic words mapped to multiple linguistic annotations. The literary genre in the *King Saud University Corpus of Classical Arabic* (henceforward referred to as KSUCCA) accounts for around 14% of the whole corpus and contains over 7 million words of poetry (42 texts); novels (2 texts); and a category entitled *Literature and Eloquence* (60 texts). We consider this to be a suitable reference set for the present study in terms of size and also register, since it exemplifies eloquent or refined language.

5.1 The Boundary Annotated Qur'ān

BAQ version 2.0 (Sawalha *et al* 2014) is used in these experiments because it includes an automatically-generated pronunciation form for each Arabic word in addition to syntactic and prosodic information in the form of part-of-speech (POS) tags and phrase break annotations. These pronunciation forms were generated in accordance with our Arabic-IPA mapping scheme and transcription technology, which uses the International Phonetic Alphabet (IPA) as standard, and outputs a phonemic transcription or citation form for each Arabic word in the text (Brierley et al. 2016; Sawalha et al. 2014). These phonemic transcriptions, together with the *tajwīd* phrase break (or punctuation) mark-up, are later used in our extraction algorithm for capturing prosodically salient words (§7). An illustrative sample from BAQ is given in Figure 3.

5.2 Pre-Processing the BAQ and KSUCCA corpora

The Qur'ān is fully vowelized but systematic treatment of short vowels and other diacritic marks is not consistent in KSUCCA due to the large number and diversity of texts. Therefore, text pre-processing to strip away all such mark up in both corpora was an important stage in order to reconcile the datasets and optimise results from KWE over the non-vowelised word forms only. This was achieved via the *Tokenizer* module in the *Standard Arabic Language Morphological*

Figure 3: Selected information tiers in BAQ for Q.70.10–11

Chapter ID	Verse ID	Word ID	MSA script	Boundary symbols		IPA phonemic transcription	Buckwalter transliteration
				Tajwīd	Major/minor		
70	10	1	وَلَا	-	-	/wala:/	walaA
70	10	2	يَسْأَلُ	-	-	/jasʔalu/	yaso>alu
70	10	3	حَمِيمٌ	-	-	/ħami:mun/	HamiymN
70	10	4	حَمِيمَاً	۞	||	/ħami:man/	HamiymaAF
70	11	1	يُبَصَّرُونَهُمْ	ۚ	||	/jubasˁsˁaru:nahum/	yubaS~aruwnahumo

Analysis (SALMA) toolkit used in this study (Sawalha 2011). When splitting characters within a word, this tokenizer imposes a constraint such that only one diacritic mark can be mapped to any given letter. Geminates are resolved by removing the *šadda* and doubling the consonant, and then mapping the first consonant in the geminate pair to *sukūn*, and the second to its short vowel. Each Arabic word is decomposed and stored as a list of (consonant, diacritic) tuples within a nested (hierarchical) data structure; and the module outputs three variant analyses: the processed vowelized word; the processed non-vowelized word, and the non-vowelized word. Figure 4 illustrates these outputs for the word form جَنَّاتٌ *janna:tun*, *gardens*, after *šadda* has been resolved (*i.e.* starting from the processed vowelized word: جَنْنَاتٌ). For KWE we use the non-vowelized word as in: جنات.

Figure 4: Decomposition of variant analyses as output from the SALMA tokenization process

		0	1	2	3	4
Processed vowelized word	جَنْنَاتٌ	ج َ	ن ْ	ن َ	ا ْ	ت ٌ
Processed non-vowelized word	جننات	ج	ن	ن	ا	ت
Non-vowelized word	جنات	ج	ن	ا		ت

We also illustrate how the tokenizer unpicks *tanwīn*, and how graphemes are mapped to IPA characters during automatic phonemic transcription of Arabic (Brierley et al. forthcoming; Brierley et al. 2016) in Figure 5. Two phonemic transcriptions are available, with or without *tanwīn*, since the final syllable will be dropped if the reciter chooses to pause after this word: /ʤanna:tun/ (with *tanwīn*) and /ʤanna:t/ (without *tanwīn*). For Arabic-IPA transcription we need to access the processed vowelised form after *šadda* has been resolved, as in: جَنْنَاتٌ.

Figure 5: Full form and pause form IPA transcriptions as outputs from our Arabic-IPA mapping algorithm

	0		1		2		3		4		5	
جَنّٰتٌ	ج	◌َ	ن	◌ْ	ن	◌َ	ا	◌ْ	ت	◌ٌ		
	ج	◌َ	ن	◌ْ	ن	◌َ	ا	◌ْ	ت	◌ُ	ن	◌ْ
one-to-one mapping	ʤ	a	n	None	n	a	aː	None	t	u	n	None
with *tanwīn*	ʤ	a	n	n			aː		t	u	n	
without *tanwīn*	ʤ	a	n	n			aː		t			

5.3 Word Counts and Output Statistics from our Experiments

After normalizing our datasets as described, the final word counts used to calculate log likelihood and extract keywords were as follows: (i) BAQ corpus: 77,430 word tokens; (ii) KSUCCA: 7,224,348; (iii) KSUCCA + BAQ as reference corpus: 7,301,778. Our approach to calculating log likelihood amalgamates the test set with the reference set to ensure that each word in the former occurs at least once in the latter. Another approach is to assume a default value for words that do not occur in the reference set. Both approaches are valid and yield the same results for a given experiment (Scott 2008). The top ten most significant nominals, verbals and grammatical words computed from our data are given in Table 2; and a more comprehensive set of nominal keywords (i.e. nouns and adjectives) with their English translation is given in Appendix I.

6. Analysis of Key Nominals in terms of Major Quranic Themes

In total, 3983 statistically significant content and function words have emerged from corpus comparison of the Qur'ān versus some seven million words from KSUCCA, with log likelihood (LL) statistics ranging from our designated cut-off point for the 99% confidence interval (>=6.63), to soaring figures such as 2328.72 for الله *Allah*, 1103.93 for ربك *your-Lord*, and 3663.35 for الذين *those-who*. We are particularly interested in nominal forms, comprising nouns and adjectives, not only because they are high in semantic content,

Table 2: Top ten nominal, verbal and function words computed via Keyword Extraction over the BAQ versus KSUCCA subset

Grammatical Category	Word	LL Statistic	Test Count	Reference Count	Test Total	Reference Total
Nominals	الله	2328.72	2153	57054	77430	7301778
	ربك	1103.93	220	537		
	السماوات	1094.65	182	233		
	عذاب	969.54	188	421		
	والأرض	748.76	157	446		
	عليم	705.50	128	229		
	ربهم	619.38	111	190		
	ربكم	545.98	102	202		
	ربنا	533.05	106	257		
	مبين	528.28	97	181		
Verbals	آمنوا	1568.08	263	349	77430	7301778
	كفروا	1124.46	189	253		
	قل	1057.98	294	1665		
	كنتم	912.72	188	507		
	ياأيها	902.13	142	143		
	كانوا	725.98	229	1663		
	يشاء	575.61	116	293		
	فقال	536.64	27	34653		
	يؤمنون	517.37	86	110		
	تعملون	496.13	83	109		
Function Words	الذين	3663.35	810	2668	77430	7301778
	إن	1358.01	966	20494		
	لكم	1231.10	337	1849		
	وما	903.84	646	13766		
	والذين	895.71	164	303		
	لهم	820.94	373	4798		
	هم	793.12	261	2046		
	أولئك	620.61	133	402		
	عليكم	567.31	164	1007		
	إنا	551.07	156	918		

but also because this group is more likely to be associated with phrase-final boundaries than any other grammatical type, and this has bearing on the second part of our study (§3.2; §7; §8). As we have said, our total word count for the Qur'ān is 77,430 tokens. Table 3 shows raw counts for all coarse-grained syntactic categories used in our Qur'ān dataset (BAQ) relative to boundary type (*i.e.* breaks versus non-breaks). These syntactic categories are also used in the *Quranic Arabic Corpus* (Dukes 2015). According to our data, 22.12% of nouns, and an impressive 57.93% of adjectives are likely to precede a phrase boundary in the Qur'ān.

Table 3: Frequencies of POS categories associated with boundary type in our Qur'ān dataset

POS Category	Breaks	Non-Breaks	Totals
Nouns	6343	22330	28673
Verbs	2199	17165	19364
Particles	19	8359	8378
Pronouns	262	7634	7896
Prepositions	277	7314	7591
Adverbs	86	1685	1771
Conjunctions	0	1471	1471
Adjectives/Nominals	1307	949	2256
Disconnected Letters	30	0	30
Totals	10523	66907	77430

The most significant nominal forms, comprising nouns and adjectives, are listed in Appendix I. Readers are reminded that LL statistics recorded in Appendix I are computed for the unmarked form (*i.e.* the Arabic word minus short vowels and diacritics) as explained in Section 5, and subsume all morphological variants resulting from different case endings (*e.g.* اللَّهُ /ʔalla:hu/ [nominative]; اللَّهَ /ʔalla:ha/ [accusative]; اللَّهِ /ʔalla:hi/ [genitive]), but *not* phrasal or cliticised variants such as والله *and-Allah*, and ربهم *their-Lord*, which appear as separate keywords. Readers will also note that phonemic pronunciation forms in the IPA character set are mapped to Arabic word forms in this section.

Having computed an extensive list of keywords and quintessentially Quranic concepts, the challenge now is to interpret them. Our proposed methodology draws on independent scholarship as an accessible standard: we will verify and interpret the highest-scoring subset

of key nominals in terms of the major Quranic themes identified by Fazlur Rahman as primary reference (Rahman 2009), though we also cite other scholarship.

Rahman defines eight high-level themes for the Quranic knowledge domain: (1) God; (2) Man as Individual; (3) Man in Society; (4) Nature; (5) Prophethood and Revelation; (6) Eschatology; (7) Satan and Evil; and (8) Emergence of the Muslim Community. We use these labels to manually categorise the subset of key nominals with LL scores descending from 2328.72 to 122.05 and appearing in Appendix I. We then examine this categorisation for words of particular interest, and discuss how keywords in any given category reflect a range of concepts within that theme. We also observe whether all eight major themes are equally well represented in our keyword collection. Most words are assigned to one theme only, with implicit recognition of possible association with alternative themes. A few words have proved difficult to categorise and these are discussed where relevant. Finally, English translations of Quranic verses quoted in full in this section are taken from Pickthall (2011).

6.1 Predominant Themes

Predominant themes in our keyword collection are: *God*; *Nature*; and *Eschatology*. Less well-represented themes are *Man in Society*, and *Emergence of the Muslim Community*; the former has four clear hyponyms: (فرعون, *Pharaoh*; لقوم, *for-a-people/tribe*; قوم, *a-people/tribe*; إسرائيل, *Israel*; and so does the latter: (عبد *slave*; أمة, *nation*; الصلاة *(the)-prayer*; الزكاة *(the)-almsgiving* or *zakat*. The theme of *Satan and Evil* also has one obvious member: الشيطان, *Satan*; but while this concept is not further reinforced by notions of whispering, machinations, treachery, and forsaking (all Satanic behaviours in the Qur'ān), our keyword list does contain one abstract denotation for *sin* in the word ضلال, *error*.

6.2 God

Allah (الله) emerges as the most significant content word in our experiments, with an LL statistic of 2328.72, and a total raw count of 2153 for the combined variants: اللَّهُ /ʔalla:hu/, اللَّهَ /ʔalla:ha/, and اللَّه /ʔalla:hi/. Phrasal and cliticised forms of the word are also retrieved: والله (*and-Allah*) and لله (*to-Allah*). Another indefinite form إله *god* is also significant.

The second most significant content word in Appendix I (LL 1103.93) is also a synonym for God in the Qur'ān: *your-Lord* (ربك)

[m.sing]. This has connotations of *lord* and *master*, but also of *caregiver* and *sustainer*. In the Qur'ān, the word most often appears within a genitive construction, and with a suffixed personal pronoun, suggesting an intimate relationship between God and man (Calderini 1994). Hence the forms ربهم (*their-lord*) [m.pl], ربكم (*your-lord*) [m.pl], ربنا (*our-lord*), ربي (*my-lord*), رب (*lord*), ربه (*his-lord*) which also appear in our keyword list. Usage of رب as an invocation is thought to be pre-Islamic (Böwering 2007). In the Qur'ān, it is the form in which the godhead first makes itself known (Q.96.1-3): '…Read: in the name of your Lord…' It is also more prevalent in the Meccan chapters (*ibid*), as in Q.106.3, where Allah is 'Lord of the Ka'ba' (Ruthven 2006: 92).

It is man's heartfelt task to comprehend God's attributes, Allah's most beautiful names (al-Jīlānī 1992: 21). Adjectival epithets retrieved in our experiments are as follows: عليم *all-knowing* (LL 705.50); رحيم *most-merciful* (LL 619.38); غفور *oft-forgiving* (LL 416.21); قدير *all-powerful* (LL 208.13); خبير *all-aware* (LL 181.98); العليم *the-all-knowing* (LL 170.01); حكيم *all-wise* (LL 168.71); بصير *all-seer* (LL 167.35); الحكيم *the-all-wise* (LL 146.67); سميع *all-hearing* (LL 110.76).

Here we note that of all the qualities attributed to God in the Qur'ān, mercy and forgiveness appear foremost in this list. We also note interrelationships between the terms retrieved, since frequent co-occurrences of the following are attested as companion words: oft-forgiving, most merciful (al-Sharif 2006:149); all-knowing, all-aware (ibid. 138); all-knowing, all-wise (ibid. 188). Furthermore, God's power قدير is exercised through manifestations of mercy (رحيم): creating, sustaining, guiding, and forgiving or returning to those who truly repent (Rahman 2009: 6–7).

6.3 Nature

There are many references to the natural world in the Qur'ān. The ordered universe is said to be 'Muslim' or surrendered to God (Rahman 2009: 65). This order, and the beauty and variety of creation, is incontrovertible evidence of God and of God's power (قدير), which also has connotations of measurement and proportion (al-Sharif 2006: 258): each aspect of creation is measured or guided by God, and the finitude of created things contrasts with the infinity and unity of God (Rahman 2009: 67). Some of our topmost keywords are: السماوات *the-heavens* (LL 1094.65); والأرض *and-the-earth* (LL 748.76); الأرض *the-earth* (LL 505.60).

However, we believe the most interesting finding in our keyword list relating to the Quranic theme of *Nature* is variants of *āya* (s.), *āyāt*

(pl.), translated as *sign* or *signs* in the English interlinear: بآياتنا (*with/by-our-signs*); آيات (*signs*); آياتنا (*our-signs*); بآيات (*with/by-signs*); لآيات (*for-signs*); آياته (*his-signs*); آية (*a-sign*); الآيات (*the-signs*); بالبينات (*with-the-clear-signs*); لآية (*for-a-sign*). Natural phenomena and processes point to God as originator and prime mover, and constitute reasons for belief (Rahman 2009: 68–73). The link between nature, reason and belief is clearly expressed in Qur'ān 13 *Al-Ra'd Thunder*, which catalogues the heavens, the sun and moon, firm hills, flowing streams, vineyards, ploughed lands, date-palms, and finally water as signs for '…*people who take thought*…' لِقَوْمٍ يَتَفَكَّرُونَ (Q.13.3) and '…*people who have sense*…' لِقَوْمٍ يَعْقِلُونَ (Q.13.4).

However, the full significance of the words آية and الآيات in the Qur'ān is that they also denote Quranic *verses* (Rahman 2009: 71–3) and are to be interpreted as God's *revelations* and God's *truth*. Again, many of these concepts congregate in Q.13.1:

> '…These are the verses [آيَاتُ] of the Scripture. That which is revealed to you from your Lord is the Truth [الحَقُّ]…'[8]

Finally, the status of Quranic verses as clearly revealed truth is present in the keyword بالبينات (LL 142.93), which Rahman translates as '[with] clear, manifest, and indubitable signs' (Rahman 2009: 73).

6.4 Eschatology

Quranic *Eschatology* is a predominant theme in the earlier, Meccan revelations (Sharaf 2012) and is well-represented in our keyword list. The fourth most significant content word in Appendix I is عذاب *punishment* (LL 969.54) and again العذاب *the-punishment* (LL 370.77). Other keywords signifying the concept of Hell are: أليم (*painful*); جهنم (*hell*); and النار (*the-fire*). For Heaven we have: جنات *gardens*; and أجر *reward*. We also have الأنهار *the-rivers*, as in the chapter entitled *Al-Bayyinah*, *The Clear Proof*.[9]

For Rahman, however, the fundamental Qur'ān teaching on الآخرة (*the-hereafter*) is that there will be a final day of reckoning, when every person is resurrected in the flesh, and shaken into self-awareness (Rahman 2009: 106). There are many compelling evocations of

[8] See also Q. 29.44 (*Al-Ankabut The Spider*): '…Allah created the heavens and the earth with truth [بالحَقِّ]. In it is indeed a sign [آيَةً] for believers…'

[9] Q.98.7–8: '…those who believe and do good works... / Their reward is with their Lord: Gardens of Eden, underneath which rivers flow, in which they dwell forever…'

the Last Judgement in the Meccan *ṣuras* (chapters), for example Q.75.3-4:

> '…Does man think that We shall not assemble his bones? / Yea, verily. Yea, We are able to restore his very fingers..;'[10]

The final judgement is represented in our keyword list via terms like: القيامة (*the-resurrections*); يومئذ (*on-that-day*); and يوم (*a-day*). Designation of people as *(un)believers* or *(dis)believers*, and as *wrongdoers* versus *those who do good works/deeds* also relates to the current theme, but is discussed under *Man as Individual* in Section 6.6.

6.5 Prophethood and Revelation

Understanding the notion of prophecy in Islam was central to Rahman's work (Moosa 2009). In our study, Mohammed (محمد LL 248.54) and Moses (موسى LL 171.05) are retrieved as significant names of prophets, but ورسوله (*and-his-messenger* LL 284.74) is the obvious top-scoring keyword associated with *Prophethood and Revelation* in Appendix I. This phrase frequently co-occurs with the word *Allah*, as in the following example from Q.4.136, which also introduces another keyword in this thematic category: الكتاب *the-book*, (LL 234.58), translated here by Pickthall as *scripture*:

> '…O you who believe! Believe in Allah and His messenger and the Scripture which He has revealed to His messenger, and the Scripture which He revealed before…'

The prophet Mohammed is also identified by the keyword نذير (*a-warner*) in the Qur'ān, especially in the earlier revelations as reminders of يَوْم الدِّين, literally *the-day of-religion* (Rahman 2009: 82). Another societal role and function is suggested by the keyword شهيد (*a-witness*), but this is harder to interpret. In matters of jurisprudence, it carries the usual legal sense of someone who witnesses the execution of a document, as in contracting fixed term debts.[11] However, God's prophets will also be summoned to bear witness against their respective communities on the last day (Rahman 2009: 31).

[10] Q.70.43–4: '…The day when they come out from the graves in haste, as if racing to a goal, / With eyes aghast, disgrace stunning them: Such is the Day which they are promised…'

[11] Q.2.282: '…And call to witness, from among your men, two witnesses. And if two men be not (at hand) then a man and two women, of such as you approve as witnesses...And the witnesses must not refuse when they are summoned…'

Furthermore, Allah is cited as a witness to the truth of revelation in altercations between Mohammed and disbelieving tribesmen.[12]

Another word considered under the present theme is a relatively high-scoring adjectival form: مبين *clear* (LL 528.28). This has the same root (*b-y-n*) as a keyword we discussed under *Nature* (§6.3): بالبينات (*with-the-clear-signs*). On inspection, it seems that the word form مبين is used in the *Qur'an* to qualify nouns such as *book* (Q5.15; Q12.1) and *warner* (Q7.184; Q11.25).

Finally, we have chosen to categorize the keywords صراط (*a-way/path*); هدى (*guidance*); and مستقيم (*straight*) primarily under *Prophethood and Revelation*, since Mohammed is one of those sent on a straight path (*i.e.* the prophets), and since Allah in his mercy has sent down the book (*i.e.* the prophetic revelation) as a guidance to mankind so that they may find God. The interplay between these ideas is beautifully present in this simple statement from Q.11.56: '…My Lord is on a straight path…'[13]

6.6 Man as Individual

The subset of keywords associated with the theme of *Man as Individual* is perhaps remarkable for what it does *not* contain. There is mention of *souls* and *hearts*: أنفسهم (*their-souls* — literally *themselves* LL 259.63); أنفسكم (*your-souls* — literally *yourselves* LL 149.14); نفس (*a-soul* LL 106.29), and قلوبهم (*their-hearts* LL 255.28), but no occurrence of concepts like *knowledge* (root علم *ʿ-l-m*) and *reason* (root عقل *ʿ-q-l*) which distinguish mankind from the rest of creation, including the angels (Rahman 2009:19).

However, keywords associated with this theme do expose basic polarities in the domains of action and belief. Hence we have: الصالحات (*the-good-deeds*) and المحسنين (*the-doers-of-good*), versus الظالمين (*the-wrongdoers*) and again الظالمون (*the-wrongdoers*). Similarly, we have: الكافرين (*the-disbelievers*) and again للكافرين (*for-the-disbelievers*), versus مؤمنين (*believers*). Words which share the root *k-f-r* (*e.g.* الكافرين *the-disbelievers*) have connotations of pride, resistance,

[12] Q.46.7–8: '…when Our clear revelations are recited to them, those who disbelieve say of the Truth when it reaches them: This is mere magic. / Or do they say: He has invented it? Say (O Muhammed): If I have invented it, still you have no power to support me against Allah…He suffices for a witness between me and you…'

[13] See also Q.36.3–6: '…You are of those sent / On a straight path, / A revelation of the Mighty, the Merciful, / That you may warn a people whose fathers were not warned, so they are heedless…'

rejection (Rahman 2009:21), and ingratitude (Ruthven 2006:93). There are also interesting contrasts in our list between words derived from the root *ḍ-l-l* (e.g. ضلال error), and words derived from *ḥ-q-q* and *ṣ-d-q*: بالحق (*by-the-truth*); الحق (*the-truth*); and صادقين (*truthful*). In the Qur'ān, the former *ḍ-l-l* signifies pointless and misguided action.[14] For Rahman, deviance is associated with impermanence. It is set against the notions of guidance and the straight path (§6.5), and also that which abides forever (§6.2) and with truth (Rahman 2009: 36).

Polarities and contrasts in the Qur'ān have long been noted. Amongst the classical exegetes who deal with 'themes' (*maʿānī*) is the Andalusian jurist Ibn Juzayy al-Kalbī (d. 741/1340) of Granada. In his *Tashīl li-ʿulūm al-tanzīl*, he argues that the Qur'ān can be thematically divided into two binary motifs: persuasion to do good (*targhīb*) and eschewing evil (*tarhīb*).[15] Ultimately, the Prophet is himself described in dual terms as a warner of impending judgement (*nadhīr*) and as a harbinger of glad tidings (*bashīr*) (Ibn Juzayy 1995: 8–9).

6.7 Hapaxes

There are three hapaxes in the keywords list. These have been corroborated and examined in context via an authoritative, indexed dictionary of the Quranic language domain (*al-Muʿjam al-Mufahras*) originally assembled by one Fu'ad Abd al-Baqi (d. 1388 AH/1968) and now made available via Al-Lahham (2007). The first is عمرا (lifetime) in Q.10.16 which is an adverb of time (*mafʿūl fihī*) from the verb لبثت (*I dwelt*) (Haywood and Nahmad 1993: 393). The second hapax is the word أحمد (*Ahmad*) in Q.61.6 which in popular Muslim lore is the name for Mohammed in heaven. Muslim missionaries such as Ahmad Deedat use the verse in polemics with Christian televangelists by arguing that أحمد is a translation of the Paraclete mentioned in John 14:16 (Deedat 1993: 59–63). The final hapax is the word أبًّا in Q.80.31 which is translated as herbage or grass (Al-Lahham 2007: 33).

[14] Q.1.6–7: '…Show us the straight path / …not the (path)...of those who go astray…'

[15] '…ponder over the Qur'ān, you will find promises and threats mentioned together…one being clarified by the other as the poet said: "By opposites things are clarified" (*fa biḍiddihā tatabayyanu l-ashyā'*)…'

7. *Madd* Extraction Algorithm

Tajwīd rules are well-defined and thus amenable to algorithmic formulation. In previous work, we have used Regular Expression (RE) search patterns over Arabic text for automatic retrieval of all verse-terminal and in-verse *qalqalah* sites in the Qur'ān (Brierley et al. 2014). Qalqalah or 'vibration' is another prosodic effect applied to a subset of Arabic consonants {ق ط د ج ب} under certain constraints during *tajwīd* recitation. For the present study, we have developed REs for capturing instances of *madd* or 'prolongation' in Quranic Arabic, as per the distinct types described in Section 3 of this paper. Input data is again drawn from the BAQ corpus but for some *madd* types this new algorithm operates over the IPA phonemic transcription tier in BAQ version 2.0 (Sawalha et al. 2014), as well as the plain text tier used for *qalqalah* capture (Brierley et al. 2014). A phonemic transcription is mapped to each fully vowelized Arabic word in version 2.0 of the corpus, complete with case endings and *tanwīn* as in: الْمَنْفُوشِ (fluffed up), rendered as the IPA string /ʔalmanfuːʃi/, and أَحْلَامٍ (dreams) rendered as the IPA string /ʔaħlaːmin/ (§3.3).

7.1 Regular Expression Search Pattern for Madd *before Pause*

In this section, we present a summary of our *madd* extraction algorithm, concentrating on abstract definition and automatic extraction of *madd* before pause (§3.1.2) since this corresponds to the device of final syllable lengthening which is in turn associated with phrase-final items and phrase edges in metrical and syntactic structure, and with informativeness (§3.2).

Arabic Natural Language Processing entails specification of each letter and diacritic mark via its unique Unicode number rather than working directly with the standard Arabic alphabet for reasons to do with interoperability; and the same applies when working with special character sets such as the IPA. Examples would be the word وَمَا represented as the Unicode string: ['u0648', 'u064E', 'u0645', 'u064E', 'u0627'][16]; and the transcription /fiː/ represented as the Unicode string: ['u0066', 'u0069', 'u003A'][17]. Hence our REs for extracting different types of *madd* specify Unicode ranges and single characters because they are designed to operate over word-level tokens (or

[16] Arabic Unicode Consortium: http://unicode.org/charts/PDF/U0600.pdf
[17] Unicode IPA Extensions: http://unicode.org/charts/PDF/U0250.pdf

alternatively tokenized phonemic transcriptions) consisting of Unicode sequences. The RE formula for identifying *madd* before pause interacts with the plain text tier only and is decomposed in Figure 6.

Figure 6: Rule-ordering in RE over Arabic Unicode for extracting *madd* before pause

u'[\u0621-\u0652]+[\u0627,\u0648,\u064A][\u0652]*[\u0621-\u064A][\u064B-\u0650]\Z'	
[\u0621-\u0652]+	at least one occurrence of any Arabic letter
[\u0627,\u0648,\u064A]	one of the *madd* letters: { ا و ي }
[\u0652]*	sukūn diacritic (optional)
[\u0621-\u064A]	any Arabic letter
[\u064B-\u0650]\Z	any short vowel or *tanwīn*: {ً ٌ ٍ َ ُ ِ} at end of string

To qualify as an instance of *madd* before pause, the word must be tagged as phrase-terminal as well as meeting all the ordered constraints in Figure 6. This is enabled by the fact that all words in the BAQ corpus are tagged as either breaks (terminals) or non-breaks, depending on whether or not they immediately precede a *tajwīd* pause mark: وَقف *waqf*, *stop*. All types of stop are considered (whether compulsory, recommended, or optional) in our coarse-grained boundary annotation scheme of break versus non-break. Boundary annotations in BAQ are described in more detail in Brierley et al. (2012); they are derived and authenticated via *tajwīd* mark-up in a reputable edition of the Qur'ān,[18] and in a widely-used recitation style: *ḥafṣ bin ʿāṣim* (cf. Sharaf 2004).

7.2 Regular Expression for Madd *Caused by Sukūn and* Šadda *before Pause*

Madd caused by *sukūn* and *šadda* requires prolongation for up to six morae and is of interest to the present study when associated with *waqf* boundary annotations. It also provides an opportunity to illustrate RE capture over IPA phonemic transcriptions. The RE is expressed in terms of *either/or* (as signified by the pipe symbol |) namely: the search is *either* for a *madd* letter /aː/, /iː/, /uː/ *or* a diphthong /aj/, /aw/. This rule is decomposed in Figure 7.

[18] http://tanzil.net/download

Figure 7: Ordering of constraints in Regular Expression over IPA phonemic transcriptions for extracting *madd* caused by *sukūn* and *šadda*

u‘*.[a,u,i]\u02D0.+.+|.*a[w,j].+.+’			
.*[a,u,i]\u02D0.+.+		.*a[w,j].+.+	
.*	optional character	.*	optional character
[a,u,i]\u02D0	a *madd* letter { ا و ي }	a[w,j]	a diphthong
.+.+	at least 2 more characters	.+.+	at least 2 more characters

It can also be seen from Figure 7 that identification of *madd* caused by *sukūn* and *šadda* requires scrutiny of two consecutive characters (**.+.+**) following the *madd* letter or diphthong. This is because we need to verify whether the next two characters are identical (*e.g.* /waɦa**ːʤʤ**ahu/, *and-argued-with-him*), or whether we are dealing with pharyngealized sounds (*e.g.* /ʔasˁsˁira**ːtˁ**a/, *(to)-the-path*). If we encounter the latter, then further rules are required to check whether the pharyngealized sound is doubled or not (*e.g.* /x**aːsˁsˁ**atan/, *exclusively*).[19] Finally, as with *madd* before pause, captured items are cross-checked against boundary annotations to finalize the frequency distribution for instances of *madd* caused by *sukūn* and *šadda* before pause.

7.3 Performing the Searches

At runtime, our algorithm converts the BAQ corpus into a unified data structure or list of lists. Inner lists within this nested structure each hold an Arabic word mapped to its chapter, verse, and word-in-verse ID, plus other linguistic information such as its part-of-speech, IPA transcription, English interlinear translation, and break type. The algorithm then iterates over the corpus line by line (*i.e.* over each inner list) checking for pattern matches in the plain text tier or IPA transcription tier for each *madd* type as defined by its unique RE. Instances of each type are then stored in separate lists as program output. These lists are later used to generate frequency distributions for each *madd* type as further output. When all this information is later combined with the results from Keyword Extraction, our quantitative analysis yields original insights into prosody as a mechanism for identifying important words and enhancing their salience. For example, the word form الْعَالَمِينَ *al-ʿalāmīna*, *the-worlds* occurs 60 times in the Qur’ān. As well as emerging as a statistically significant keyword

[19] We use the symbol /x/ rather than /χ/ to transcribe the Arabic letter خ (Brierley et al. 2016).

(LL 203.27), this variant always occurs at phrase edges and so always features *madd* before pause. Hence the accusative case mark would always be dropped in practice and its corresponding phonemic transcription would therefore be: /ʔalʕaːlamiːn/.

8. Discussion of Results

In this section, we compare statistics from Keyword Extraction (KWE) and extraction of *madd* before pause as quantitative evidence to address the following research questions:

1. *Is there evidence that statistically significant words in the Qur'ān are also made prosodically prominent, via tajwīd annotation of increased duration in the syllable before pause in their variant realisations? If so, how often does this occur?*
2. *Which statistically significant words in the Qur'ān also occupy prime (phrase-final) position in terms of metrical and syntactic structure with unusual frequency?*

From our KWE experiments, we can access the raw frequencies of unmarked keywords with high LL values in the Qur'ān versus KSUCCA. These raw frequency totals represent the combined counts for all fully vowelized morphological variants subsumed under any given unmarked form. For example, السماوات (*al-samāwāt*, *the heavens*) has a combined count of 182, and subsumes two variants: السَّمَاوَاتِ (*al-samāwāti*) *and* السَّمَاوَاتُ (*al-samāwātu*). The former occurs 175 times in the Qur'ān, whereas the latter only occurs 7 times. If the former is frequently associated with pre-pausal lengthening in *tajwīd* annotation, then that is exactly the sort of evidence we seek in relation to our research questions above.

It transpires that there is only one instance of السَّمَاوَاتِ marked up with *madd* before pause in our Qur'ān dataset. However, in Section 7.1, we drew attention to the excellent example of الْعَالَمِينَ (*al-ʿalāmīna*, *the worlds*) which occurs 60 times in the Qur'ān. This also represents the total count for the unmarked form and keyword: العالمين. So as well as accounting for a statistically significant lexical item (العالمين LL 203.27), this variant is always annotated with *madd* before pause because it *always* occurs at phrase edges; its phrase-final position is not accidental since it is determined by syntax and sentence/utterance planning. Hence, from this fusion at the prosody-syntax-semantics interface, we deduce that this word, and its denotation, represented by a unique sound sequence and rhythmic pattern, is of particular

importance in the *Qur'ān* as God's message to mankind: there is/are (an) unseen world(s) beyond the concrete world mankind inhabits.[20]

8.1 How often are Keywords in the Qur'ān *made Prosodically Prominent?*

From our corpus comparison experiments in Section 5, we have extracted a total of 3983 keywords with an LL statistic >= 6.63, the cut-off for statistical significance according to Rayson (2008). We now find that 53.73% of these are tagged as either nouns or adjectives: ((2140/3983)*100). These are what we have termed *nominal keywords*; and these are of interest to this study not only because they are high in semantic content, but also because they are more likely to terminate a phrase than any other grammatical type. Statistics presented in Table 3 (§6) reveal that (based on raw frequency) adjectives are most likely to be phrase-final, co-occurring with boundaries 58% of the time; while nouns co-occur with boundaries 22% of the time; and nouns-plus-adjectives co-occur with boundaries 25% of the time. In contrast, verbs co-occur with boundaries 11% of the time; adverbs co-occur with boundaries 5% of the time; and verbs-plus-adverbs co-occur with boundaries 11% of the time in our data.

Table 4: Co-occurrence statistics of grammatical types with punctuation (pause marks) in the *Qur'ān*

Breaks	Non-Breaks	% Co-Occurrence with Boundaries
1307 Adjectives	949 Adjectives	(1307/2256)*100 = **58%**
6343 Nouns	22330 Nouns	(6343/28673)*100 = **22%**
1307 Adjectives + 6343 Nouns	949 Adjectives + 22330 Nouns	(7650/30929)*100 = **25%**
2199 Verbs	17165 Verbs	(2199/19364)*100 = **11%**
86 Adverbs	1685 Adverbs	(86/1771)*100 = **5%**
2199 Verbs + 86 Adverbs	17165 Verbs + 1685 Adverbs	(2285/21135)*100 = **11%**

Turning to prosodically prominent items, our madd extraction algorithm retrieves 5412 instances of madd before pause in our Qur'ān

[20] This same concept is present in the Christian (Nicene) creed: '…Credo in unum Deum, Patrem *omnipotentem, factorem caeli et terrae,* ***visibilium et invisibilium****…*'

dataset. Further manual and statistical analysis reveals that 71.36% of these are tagged as either nouns (total 3036) or adjectives (total 826); and that the set of noun-plus-adjective types annotated with madd before pause comprises 1215 items.

Finally, considering the intersection of nominal keywords and noun-adjective types with madd before pause, we find 526 common items. This gives a percentage crossover of 43.29%: ((526/1215)*100). There is therefore considerable evidence of target prosodic highlighting (*i.e. tajwīd* annotation of *madd* before pause) associated with statistically significant keywords in the Qur'ān.

8.2 Which Key Quranic Concepts are Most Likely to be Made Prosodically Prominent?

In this section, we consider the overlap between the subset of nominal keywords presented in Appendix I (*i.e.* keywords with higher LL figures) and their counterparts in our extracted subset of prosodically prominent nominals (i.e. nouns and adjectives) annotated with *madd* before pause. This set-overlap measure, which simply counts words common to both lists, does not account for rankings of words within these lists as in *rank-biased overlap* (Webber, Moffat and Zobel 2010) or *weighted overlap* (Pilehvar and Navigli 2015), though we do present a simple ranked ordering of items in Table 5 (§8.5).

The criteria informing interpretation of results are: (i) high LL score for unmarked form; (ii) high count for fully-vowelized variant in relation to raw count for unmarked form; (iii) good proportion of fully vowelized variant marked up with *madd* before pause relative to total count for fully vowelized form. To illustrate the application of these criteria, we first discuss the items عَلِيمٌ(*ʿalīmun*) and عَلِيمٍ(*ʿalīmin*). The total count for the unmarked form عليم (*all-knowing*) is 128, and it has a very significant LL score of 705.50. This total is made up of raw counts for all fully-vowelized variants of عليم, namely: عَلِيمٌ (100), عَلِيمٍ (6), and عَلِيمًا (22). It transpires that the word form عَلِيمًا (*ʿalīman*) is never annotated with *madd* before pause so we can assume that it is never phrase-final and that it is always pronounced with *tanwīn* during recitation. Contrast this with the variant عَلِيمٌ which co-occurs with a boundary 56% of the time (*i.e.* 56/100 instances annotated with *madd*), and the variant عَلِيمٍ which co-occurs with a boundary 100% of the time (*i.e.* 6/6 instances annotated with *madd*). In phrase-final position, both variants will be realised without *tanwīn* and will sound identical: /ʕalīm/.

8.3 Laplace Point Estimation

We now direct the reader to Appendix II, which summarizes the above data for عليم and all other matches in our keyword and madd-before-pause subsets. An important aspect of our analysis here is inclusion of a *point estimate* statistic which represents the probability that a given word (represented by the unmarked form) will be annotated with *madd* before pause (and hence will complete a phrase) in the Qur'ān. This probability seems somewhat difficult to establish since it is only fully vowelised instances that are realised with *madd*; the total count for the unmarked form subsumes raw counts for its associated morphological variants.

However, the probability that a statistically significant word (denoted by the unmarked form) will be annotated with *madd* before pause in our data can be construed more simply as a task akin to that of measuring successful rates of completion in usability studies. In such a scenario, successful task completion rate is most commonly computed as a single value or point estimate given by the formula x / n, namely: the number of successful attempts x divided by the total number of attempts n. This is known as the Maximum Likelihood Estimate (MLE).

It is also important to consider alternative methods that have been devised to compensate for extreme outcomes particularly when sample sizes are small (Lewis and Sauro 2006). This has bearing on our data when, for example, low frequency variants such as عَليم always exhibit *madd* before pause (§8.3). Therefore, instead of simply dividing the number of successes by the number of trials (*i.e.* MLE), we implement an adjustment known as the Laplace point estimate which specifically targets extreme outcomes (*e.g.* 100% success rate) over small sample sizes. This method adds one to the numerator and two to the denominator, and is expressed as $((x+1) / (n+2))$, where x represents the number of successes, and n is the number of trials. For example, we find that the keyword عليم (*all-knowing*) is made prosodically prominent 49% of the time by rounding up the Laplace Point Estimate of 0.4846, and expressing it as a percentage. Data used in this calculation includes: (i) total count for the unmarked form (in this case 128); and (ii) number of occurrences of variant(s) with *madd* (in this case 56 + 6 = 62). Implementing Laplace, we therefore have (62+1 successes) / (128+2 trials), resulting in a probability of 49%.

8.4 Interpretation of Results

Semantically weighted Quranic words/concepts that are both 'key' (*i.e.* significant statistically) and also highly likely to be made prosodically prominent in phrase-final position can perhaps be grouped as follows, where percentages are derived from the Laplace Point Estimate score for a given item in Appendix II.

8.4.1 God

The first group concerns the nature of God, namely the attributes of Allah, Lord of the Worlds. It includes: الحكيم (*the-all-wise*) 86%; قدير (*all-powerful*) 82%; العليم (*the-all-knowing*) 79%; رحيم (*merciful*) 74%; حكيم (*wise*) 58%; بصير (*all-seeing*) 48%; خبير (*all-aware*) 41%.

8.4.2 Eschatology – and Man as Individual

The second group concerns the nature of man, manifest in the distinction between good and evil. It includes the following oppositions:المحسنين (*the-doers-of-good*) 94% versus الظالمين (*the-wrongdoers*) 76%, and الظالمون (*the-wrongdoers*) 65%; مؤمنين (*believers*) 94% versus الكافرين (*the-disbelievers*) 57%; المتقين (*the-god-fearing*) 60% versus المجرمين (*the-criminals*) 62%. It may also include the masculine plural form: خالدون (*abiding-forever*) 96%, either in Paradise or Hell. Such polarisation is characteristic of the Quranic theme of Eschatology, and of the Qur'ān more generally (§6.6).

8.4.3 Eschatology – and Prophethood and Revelation

Finally, there is a group of powerful *end-of-the-world* concepts reminiscent of the early Meccan *ṣuras*. It includes punishment for unredeemed sin: الجحيم (*the-hellfire*) 72%; أليم (*painful*) 69%; and النار (*the-fire*) 46%. This may be juxtaposed with مستقيم (*straight*) 77%, and مبين (*clear*) 86%, defining the guidance offered by God's 'messengers', and linking the Quranic themes of Eschatology and Prophethood and Revelation.

8.5 Ranking Keyness with Respect to Madd-ness

As stated in Section 8.3, we have focussed this analysis on set overlap between the keyword nominals presented in Appendix I, and the nominals annotated with *madd* before pause presented in Appendix II. In Appendix III, we have ranked the 43 items common to both lists based on their Laplace Point Estimate score mapped to corresponding rank in terms of Log Likelihood. Items with the most

congruent rankings are presented in Table 5 and again reflect the Quranic themes of *God*, *Nature*, and *Eschatology*, with a notable exception being إسرائيل (*Israel*).

Table 5: Quranic words/concepts with similar rankings for keyness and ominence

La Place Ranking	Unmarked Form	LL Rank	Literal English Translation	Quranic Theme
12	الظالمين	10	the-wrongdoers	Eschatology
23	بصير	27	(all)-seeing	God
24	النار	29	the-fire	Eschatology
25	خبير	21	(all)-aware	God
29	إسرائيل	30	Israel	Man in Society
31	بالبينات	35	with-the-clear-signs	Nature
32	الآيات	33	the-signs	Nature
33	ضلال	28	error	Satan/Evil

8.6 Interaction of Keyness and Madd *caused by* Sukūn *and* Šadda *before Pause*

In Section 7.2, we presented a second RE search pattern to capture instances of *madd* caused by *sukūn* and *šadda* before pause. While these do not yield high LL statistics, we have assembled the highest-scoring instances in descending order of Laplace Point Estimate in Table 6.

Table 6: Top ten words annotated with *madd* caused by *sukūn* and *šadda* before pause, and ranked in descending order of Laplace Point Estimate score, representing the probability that a given word will be made prosodically prominent in the Qur'ān

Rank	Laplace Score	Unmarked Form	Literal English Translation	Marked Form	Raw Counts		
					Total Count: Unmarked Form	Freq. Marked Variant	Freq. Marked Variant with *Madd*
1	**0.875**	الضالين	the-astray	الضَّآلِّينَ	6	6	6
2	**0.8571**	كافة	all-(mankind)	كَآفَّةً	5	5	5
3	**0.8571**	جان	jinn	جَانٌّ	5	5	5

4	**0.8125**	دابة	a-moving-creature	دَابَّة	14	12	12
5	**0.80**	الضالون	the-astray-ones	الضَّالُّونَ	3	3	3
6	**0.80**	الحاقة	the-inevitable-reality	الْحَاقَّةُ	3	3	3
7	**0.75**	صَافَّاتٍ	spreading-(their-wings)	صافات	2	2	2
8	**0.75**	لضالون	surely-have-gone-astray	لَضَالُّونَ	2	2	2
9	**0.75**	يتماسا	they-both-touch-each-other	يَتَمَاسَّا	2	2	2
10	**0.75**	ضالين	astray	ضَالِّينَ	2	2	2

These are of interest to this study since they again exhibit a correlation between keyness and prosodic prominence. Several are associated with the Quranic theme of Eschatology, and with deviance from the straight path[21]: الضالين (88%), الضالون (80%), لضالون (75%), ضالين (75%).

9. Conclusions and Further Work

Our first contribution in this paper is a set of Quranic keywords extracted via corpus comparison of the Qur'ān versus a 7 million-word reference set of Classical, literary Arabic. These keywords represent a particular notion of linguistic salience, namely: statistical significance. We have then verified the meaningfulness of nominal keywords (*i.e.* nouns and adjectives) in this set by qualitative, interpretative techniques, mapping them to major Quranic themes in Islamic scholarship. The themes of *God*, *Nature*, and *Eschatology* are well represented in our results; and some of our most interesting findings, and top-scoring keywords, emphasise: God's mercy and forgiveness (*God*); natural phenomena as 'signs' or proof of God (*Nature*); and the reality of heaven, hell, and a final day of reckoning (*Eschatology*).

We then hypothesise that keywords extracted via computation over the Qur'ān may already be encoded as meaningful in rules governing traditional Quranic recitation, or *tajwīd*. Our specific focus is the *tajwīd* sub-category of *madd* (prolongation) before pause. According to this rule, words in phrase-final position are made perceptually prominent via abnormal prolongation of Arabic long vowels and diphthongs when they occur in the syllable before pause. Pre-boundary lengthening has long been accepted as a cue to prominence, and

[21] This concept of disorientation and losing one's way is also very reminiscent of Dante's *Inferno* (Canto 1.1–3): '…Nell mezzo del cammin di nostra vita / mi ritrovai per una selva oscura / ché la diritta via era smarrita…'

prosodic prominence and phrasing are widely associated with informativeness as well as signifying metrical structure. Thus our second contribution is to the Islamic Studies sub-field of *tajwīd*: we interpret the phenomenon of *madd* before pause as final syllable lengthening, and even prosodic boundary mark-up, and so link the linguistic insights informing *tajwīd* with modern linguistic theory.

Another important aspect of this work is to ascertain whether statistically significant words in the Qur'ān are also highlighted prosodically via the device of *madd* before pause. This is original research since a potential correlation between distributionally significant and prosodically prominent words is hitherto unexplored both in the Qur'ān and in Corpus Linguistics more generally.

To address this novel research question, we have made a further contribution in the form of our bespoke *madd* extraction algorithm. This is designed to operate over automatically-generated phonemic transcriptions of Arabic words in our *Boundary-Annotated Qur'ān*, as well as over plain text, and defines a set of constraints or regular expressions to capture all instances within each sub-category of *madd* in our dataset.

We have found considerable evidence of nominal keywords aligned with phrase edges and bearing the hallmarks of pre-boundary lengthening. In our data, this correlation occurs 43.29% of the time, since 526 keywords also appear in the extracted set of nouns and adjectives tagged with *madd* before pause: ((526/1215)*100). We have also found convincing evidence that certain Quranic keywords are also highly likely to be reserved for phrase-final position and to be made prosodically prominent with extra prolongation (*madd*) before pause.

This probability is somewhat difficult to establish because our keywords are computed from the raw frequencies of unmarked forms in the Qur'ān, whereas all instances of *madd* before pause are marked. A notable and straightforward example is the keyword العالمين with a log likelihood score of 203.27. This occurs 60 times in the Qur'ān and is *always* realised via the same vowelization (الْعَالَمِينَ, al-ʿalāmīna, the worlds) *and* with *madd* before pause.

To compute the majority of less straightforward cases, we have used the Laplace Point Estimate statistic. This serves two purposes: (i) it represents the number of successes (i.e. instances of *madd* before pause) relative to the number of trials (i.e. total raw frequency) for any given item as a single value which is easy to interpret; and (ii) it compensates for extreme outcomes (*e.g.* 100% success rate) over small sample sizes.

Using this metric, we have again found that keywords representing the themes of *God*, *Nature* and *Eschatology* are more likely to emerge as semantically weighted in terms of distributional and prosodic significance. Top-scoring examples are: العالمين (*the-worlds*), with *madd* before pause 98% of the time; صادقين (*truthful* [m.pl]), with *madd* before pause 97% of the time; خالدون (abiding-forever), with *madd* before pause 96% of the time. Furthermore, examining congruent rankings of keyness with respect to *madd*-ness in a list of 43 items ranked in descending order of Laplace score, the same themes emerge, along with a notable exception: إسرائيل (*Israel*).

To sum up, we have found evidence that distributionally significant, and quintessentially Quranic keywords also occur strategically in phrase-final position so as to maximise their prominence, and thus meaningfulness, for reader, reciter, and aural recipient. Further work is needed to determine the extent of this congruence between keyness and prominence in the Qur'ān, and to determine whether such congruence is peculiar to this text, or a feature of Arabic more generally, and/or more universally for other languages. Further work in cross-referencing and interpreting concepts and annotated phenomena in traditional Arabic linguistics in the light of contemporary linguistic theory is also needed.

Addresses for correspondence: C.Brierley@leeds.ac.uk

REFERENCES

Abdul-Fattah, A. 1989. *Tajwīd-Ul-Qur'ān: A New Approach to Mastering the Art of Reciting the Holy Qur'ān.* (London)

Al-Ḍabbāʿ, A. 1997. *Minḥat dhī l-Jalāl fi sharḥ Tuḥfat l-Atfāl.* (Riyadh)

Al-Jīlānī, H.A. 1992. *The Secret of Secrets*. (Cambridge)

Al-Lahham, M.S. 2007. *Al-Mu'jam al-Mufahras li-alfāẓi l-Qur'āni l-Karīm.* (Beirut)

Al-Masīrī, K. 2002. *Al-Jāmi'fi qirā'āt al-Qur'ān al-Karīm* (Compendium of Quranic Tajwīd. Cairo)

Ali, A.Y. 2000. *The Holy Qur'ān*: Translated by Abdullah Yusuf Ali. (Ware, Herts)

Alrabiah, M., A. Al-Salman and E.S. Atwell. 2013, 'The Design and Construction of the 50 Million Words KSUCCA', in *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics*, University of Leeds. 5–8

Al-Sharif, M.M. 2006. *Asmā' Allah Al-Ḥusnā* (Allah's Most Beautiful Names). (Beirut)

Atwell E., J. Dickins and C. Brierley. 2013. *Natural Language Processing Working Together with Arabic and Islamic Studies*. EPSRC [EP/K015206/1]

Bohas, G., J.P. Guillaume and D.E. Kouloughli. 1990. *The Arabic Linguistic Tradition*. (London)

Böwering, G. 2007. 'God and His Attributes', in *Encyclopaedia of the Qur'ān*. (Leiden)

Brierley, C., M. Sawalha and H. El-Farahaty, (Forthcoming). 'Translating Sacred Sounds: Encoding tajwīd Rules in Automatically-generated IPA Transcriptions of Quranic Arabic'. Chapter submission for the *Routledge Handbook of Arabic Translation*

Brierley, C., M. Sawalha, B. Heselwood and E. Atwell. 2016. 'A Verified Arabic-IPA Mapping for Arabic Transcription Technology, Informed by Quranic Recitation, Traditional Arabic Linguistics, and Modern Phonetics', *JSS* 61:1, 157–86

Brierley, C., M. Sawalha and E. Atwell. 2014, 'Tools for Arabic Natural Language Processing: a Case Study in *qalqalah* Prosody, in *Proceedings of Language Resources and Evaluation Conference (LREC) 2014*, Reykjavik, Iceland. 283–7

—— 2012. 'Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing', in *Proceedings of LREC 2012*, Istanbul, Turkey, 1011–16

Brierley, C. and E. Atwell. 2007. 'Prosodic Phrase Break Prediction: Problems in the Evaluation of Models against a Gold Standard, *Traitement Automatique des Langues* 48:1

Calderini, S. 1994. Quranic Exegesis of ʿalamin, in rabb al-ʿalamin, *Bulletin of the School of Oriental and African Studies* (BSOAS) 57:1. 52–8

Calhoun, S. 2010a. 'Centrality of Metrical Structure in Signalling Information Structure: A Probabilistic Perspective, *Language* 86:1, 1–42

—— 2010b. 'How does Informativeness Affect Prosodic Prominence?', *Language and Cognitive Processes*, 25:7–9, 1099–1140

Dar-Al-Maarifah. 2008. *Tajweed Qur'ān: With Meaning Translation in English and Transliteration*. (Damascus)

Deedat, A. 1993. *The Choice: Islam and Christianity, Volume 1*. (Saudi Arabia)

De Jong, K. and B. Zawaydeh, B. 2002. 'Comparing Stress, Lexical Focus, and Segmental Focus: Patterns of Variation in Arabic Vowel Duration', *Journal of Phonetics*, 30:1, 53–75

—— 1999. 'Stress, Duration, and Intonation in Arabic Word-Level Prosody', *Journal of Phonetics*, 27:1, 3–22

De Jong, K.J. 1995. 'The Supraglottal Articulation of Prominence in English: Linguistic Stress as Localized Hyperarticulation', *Journal of the Acoustical Society of America*, 97:1, 491–504

Dukes, K. 2015. 'Statistical Parsing by Machine Learning from a Classical Arabic Treebank'. Ph.D. thesis, University of Leeds

Frota, S., J. Butler, S. Correia, C. Severino and M. Vigário. 2012. 'Pitch First, Stress Next? Prosodic Effects on Word Learning in an Intonation Language', in *Proceedings of the 36th Annual Boston University Conference on Language Development*. 190–201

Fry, D.B. 1958. 'Experiments in the Perception of Stress', *Language and Speech* 1, 120–52

Ghawthānī, Y.A. 1996. *ʿIlm al-Tajwīd*: (Science of Tajwīd). (Jeddah)

Gordon, M. and P. Munro. 2007. 'A Phonetic Study of Final Vowel Lengthening in Chickasaw 1', *International Journal of American Linguistics* 73:3, 293–330

Gordon, M. 2014. 'Disentangling Stress and Pitch Accent: A Typology of Prominence at Different Prosodic Levels', in H. van der Hulst (ed.), *Word Stress: Theoretical and Typological Issues*. (Oxford). 83–118

Gundel, J.K. 1988. 'Universals of Topic-Comment Structure', in M. Hammond, E. Moravczik and J. Wirth (eds), *Studies in Syntactic Typology* (Amsterdam). 209–39

Halliday, M.A. 1967. 'Notes on Transitivity and Theme in English: Part 2', *Journal of Linguistics* 3:2, 199–244

Haywood, J.A. and H.A. Nahmad. 1993. *A New Arabic Grammar of the Written Language*. (London)

Hellmuth, S. 2007. 'The Relationship between Prosodic Structure and Pitch Accent Distribution: Evidence from Egyptian Arabic', *The Linguistic Review*, 24:2–3, 291–316

—— 2014. 'Investigating Variation in Arabic Intonation: the Case for a Multi-Level Corpus Approach', in S Farwaneh and H Ouali (eds), *Perspectives on Arabic Linguistics XX1V–XXV* (Studies in Arabic Linguistics, vol. 1, Amsterdam). 63–89

Heselwood, B. and Z.M. Hassan. 2011. 'Introduction', in Z.M. Hassan and B. Heselwood (eds), *Instrumental Studies in Arabic Phonetics* (Amsterdam)

Ibn Juzayy, M. 1995. *Al-Tashīl li-ʿUlūm al-Tanzīl*. Vol. I. (Beirut)

Jun, S. 2005. 'Prosodic Typology', in S. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. (Oxford). 430–58

Klatt, D.H. 1976. 'Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence', *The Journal of the Acoustical Society of America* 59:5, 1208

Ladd, D.R. 1996. *Intonational phonology*. (Cambridge Studies in Linguistics 79. Cambridge)

Lewis, J.R. and J. Sauro. 2006. 'When 100% Really Isn't 100%: Improving the Accuracy of Small-Sample Estimates of Completion Rates', *Journal of Usability Studies* 3:1, 136–50

Liberman M.Y. and K.W. Church. 1992. 'Text Analysis and Word Pronunciation in Text-to-Speech Synthesis', in S. Furui and M.M. Sondhi (eds), *Advances in Speech Signal Processing* (New York)

Muhammad, A.B. 2012. 'Annotation of Conceptual Co-reference and Text Mining the Qur'an'. Ph.D. thesis, University of Leeds

Nespor, M., M. Shukla and J. Mehler. 2011. 'Stress-timed vs. Syllable-timed Languages', in M. van Oostendorp, C.J. Ewen, E. Hume and K. Rice (eds), *The Blackwell Companion to Phonology (II)* (Oxford). 1147–59

Pickthall, M.M. 2011. *The Meaning of the Glorious Qur'ān: An Explanatory Translation*. (Birmingham)

Pilehvar, M.T. and R. Navigli. 2015. 'From Senses to Texts: An All-in-one Graph-based Approach for Measuring Semantic Similarity', *Artificial Intelligence* 228, 95–128

Rahman, F. 2009. *Major Themes of the Qur'an*[2]. (Chicago)

Rayson, P. 2008. 'From Key Words to Key Semantic Domains', *International Journal of Corpus Linguistics* 13:4, 519–49

Rayson, P. and R. Garside. 2000, 'Comparing Corpora using Frequency Profiling', in *Proceedings of the Workshop on Comparing Corpora* (Association for Computational Linguistics). 1–6

Read, I. and S. Cox. 2007. 'Stochastic and Syntactic Techniques for Predicting Phrase Breaks', *Computer Speech & Language*, 21:3, 519–42

Read with Tajweed. 2016. Online. Accessed: 23.11.2016. http://www.readwithtajweed.com/

Ruthven, M. 2006. *Islam in the World*. (London)

Ryding, K.C. 2014. *Arabic: A Linguistic Introduction*. (Cambridge)

Sawalha, M., C. Brierley and E. Atwell. 2014. 'Automatically Generated, Phonemic Arabic-IPA Pronunciation Tiers for the *Boundary Annotated Qurān Dataset for*

Machine Learning (version 2.0)', in *Proceedings of the 2nd Workshop for Language Resources and Evaluation of Religious Texts, LREC 2014* (Reykyavik). 42–7

Sawalha, M. 2011. 'Open-Source Resources and Standards for Arabic Word Structure Analysis: Fine-Grained Morphological Analysis of Arabic Text Corpora'. Ph.D. thesis, University of Leeds

Scott, M. 2008. *Corpus Linguistics Summer Institute*. University of Liverpool. June-July 2008.

—— 1997. 'PC Analysis of Key Words - and Key Key Words', *System* 25:2, 233–45

Selkirk, E.O. 2005. 'Comments on Intonational Phrasing in English', in S. Frota, M. Vigário and M.J. Freitas (eds), *Prosodies* (*with Special Reference to Iberian Languages*) (Berlin and New York). 11–58

—— 1986. ‚On Derived Domains in Sentence Phonology', in *Phonology Yearbook* 3, 371–405

Shah, M. 2003. 'Exploring the Genesis of Early Arabic Linguistic Thought: Qur'anic Readers and Grammarians of the Kūfan Tradition (Part I)/تطور الدراسات اللغوية بين القراء والنحاة الكوفيين (القسم الأول), *Journal of Qur'anic Studies* 5:1, 47–78

Shaikh, S. and Khatri, K. 2014. *eMuslim*. Online. Accessed: 30.06.2014. http://www.emuslim.com/Quran/Translation_English.asp

Sharaf, A.B. 2012. 'Annotation of Conceptual Co-reference and Text Mining the Qur'ān'. Ph.D. thesis, University of Leeds

Sharaf, J.A.M. 2004. مصحف الصحابة في القراءات العشر المتواترة من طريق الشاطبية والدرة *mushaf as-sahabah fi al-qira'at al-'ashr al-mutawatirah min tariq ash-shatibyyah wa al-durrah*. (Tanta)

Taylor P. and A.W. Black. 1998. 'Assigning Phrase Breaks from Part-of-Speech Sequences', *Computer Speech and Language* 12:2, 99–117

Watson, J.C.E. 2011. 'Word Stress in Arabic', in M.V. Oostendorp, C. Ewen, E. Hume and K. Rice, (eds) *The Blackwell Companion to Phonology* (Oxford). 2990–3019

Webber, W., A. Moffat and J. Zobel, 2010. 'A Similarity Measure for Indefinite Rankings', *ACM Transactions on Information Systems (TOIS)* 28:4, 20

Wichman, A., N. Dehé and D. Barth-Weingarten. 2009. ‚Where Prosody Meets Pragmatics: Research at the Interface', in D. Barth-Weingarten, N. Dehé and A. Wichmann (eds), *Where Prosody Meets Pragmatics* (Studies in Pragmatics 8, Bingley). 1–20

Wightman, C.W., S. Shattuck-Hufnagel, M. Ostendorf and P.J. Price. 1992. 'Segmental Durations in the Vicinity of Prosodic Phrase Boundaries', *The Journal of the Acoustical Society of America*, 91:3, 1707–17

Appendix I: Nominal keywords linked to major Quranic themes

Quranic Themes (Rahman 2009): 1 - God; 2 - Man as Individual; 3 - Man in Society; 4 - Nature; 5 - Prophethood and Revelation; 6 - Eschatology; 7 - Satan and Evil; 8 - The Muslim Community; Hapax – keywords that only occur once in the Qur'ān

LL Statistic	Unmarked Word Form	Quranic Theme
2328.72	الله	1
1103.93	ربك	1
1094.65	السماوات	4
969.54	عذاب	6
748.76	والأرض	4
705.50	عليم	1
619.38	ربهم	1
545.98	ربكم	1
533.05	ربنا	1
528.28	مبين	5
505.60	الأرض	4
433.11	رحيم	1
416.21	غفور	1
370.77	العذاب	6
369.05	ربي	1
356.42	جنات	6
356.05	بآياتنا	4
354.65	أليم	6
343.11	جهنم	6
339.18	عبد	8
335.46	الصالحات	2
329.68	إله	1
324.49	بالله	1
316.07	الظالمين	2
310.43	أبي	3

LL Statistic	Unmarked Word Form	Quranic Theme
296.15	سبيل	5
295.52	بالحق	2
290.26	ابن	3
284.74	ورسوله	5
278.60	خالدين	6
276.75	الكافرين	2
269.70	فرعون	3
268.35	رب	1
259.63	أنفسهم	2
255.28	قلوبهم	2
248.54	محمد	5
234.92	الكتاب	5
234.81	آيات	4
224.10	ربه	1
221.99	القيامة	6
221.66	والله	1
215.41	السماء	4
213.77	الأنهار	6
210.11	أجر	6
208.13	قدير	1
203.49	يومئذ	6
203.27	العالمين	4
201.97	لقوم	3
200.71	صراط	5
200.30	الآخرة	6

LL Statistic	Unmarked Word Form	Quranic Theme
199.62	آياتنا	4
197.08	الحق	2
194.57	بآيات	4
191.73	مؤمنين	2
189.99	يوم	6
186.39	آلاء	4/6
185.58	خير	5
181.98	خبير	1
179.70	صادقين	2
179.31	لآيات	4
174.67	الشيطان	7
174.31	لله	1
173.80	نذير	5
171.05	موسى	5
170.01	العليم	1
168.71	حكيم	1
167.35	بصير	1
163.33	هدى	5
160.90	آياته	4
159.24	قوم	3
157.82	ضلال	7
156.73	النار	6
153.88	إسرائيل	3
152.96	آية	4
149.14	أنفسكم	2
146.67	الحكيم	1
146.37	عظيم	U
144.49	الآيات	4
144.42	مستقيم	5
142.93	بالبينات	4
141.32	الحياة	2

LL Statistic	Unmarked Word Form	Quranic Theme
140.78	خالدون	1
140.03	شهيد	5
140.00	للكافرين	2
139.26	المحسنين	2
133.10	الظالمون	2
129.55	أمة	8
127.45	أعمالهم	2
125.64	الملائكة	4/6
122.05	عمر	2 Hapax
121.01	أحمد	6 Hapax
120.51	الصلاة	8
120.15	ويوم	6
119.85	يابني	3
119.85	لآية	4
119.59	الأولين	3
117.86	للناس	3
117.13	بالآخرة	6
116.18	المتقين	6
112.88	مسمى	6
111.68	الجحيم	6
111.23	الجنة	6
110.99	قوله	5
110.79	الزكاة	8
110.76	سميع	1
107.01	المؤمنون	6
106.29	نفس	2/6
105.61	أبا	4 Hapax
105.33	الدنيا	4
104.72	المجرمين	6
104.67	المرسلين	5

Appendix II: Analysis of madd before pause on nominal keywords

Quranic Themes (Rahman 2009): 1 - God; 2 - Man as Individual; 3 - Man in Society; 4 - Nature; 5 - Prophethood and Revelation; 6 - Eschatology; 7 - Satan and Evil; 8 - The Muslim Community

Morph. Variant	Raw Count: Variant	Raw Count: Variant with Madd	Unmarked Form	Total Count: Unmarked Form	LL	LaPlace Point Estimate	Interlinear	Quranic Theme
الْعَالَمِينَ	60	60	العالمين	60	203.27	**0.9839**	the-worlds	4
صَادِقِينَ	31	31	صادقين	31	179.7	**0.9697**	[those-who-are]-truthful	6
خَالِدُونَ	24	24	خالدون	24	140.78	**0.9615**	abiding-forever	1
مُؤْمِنِينَ	33	33	مؤمنين	34	191.73	**0.9444**	believers	6
الْمُحْسِنِينَ	30	29	المحسنين	30	139.26	**0.9375**	(to)-the-good-doers	6
الْحَكِيمُ	33	28	الحكيم	42	146.67	**0.8636**	all-wise	1
الْحَكِيمِ	9	9						
مُبِينٌ	45	45	مبين	97	528.28	**0.8586**	clear	5
مُبِينٍ	39	39						
الْمُرْسَلِينَ	24	21	المرسلين	24	104.67	**0.8462**	the-messengers	5
قَدِيرٌ	37	36	قدير	43	208.13	**0.8222**	all-powerful	1

Morph. Variant	Raw Count: Variant	Raw Count: Variant with Madd	Unmarked Form	Total Count: Unmarked Form	LL	LaPlace Point Estimate	Interlinear	Quranic Theme
الْعَلِيمُ	28	22	العليم	32	170.01	**0.7941**	the-all-knower	1
الْعَلِيمِ	4	4						
مُسْتَقِيمٍ	20	19	مستقيم	32	144.42	**0.7647**	(the)-straight-path	5
مُسْتَقِيمٌ	6	6						
الظَّالِمِينَ	64	49	الظالمين	64	316.07	**0.7576**	wrongdoers	6
رَحِيمٌ	59	58	رحيم	81	433.11	**0.7349**	(the)-most-merciful	1
رَحِيمٍ	2	2						
الْجَحِيمِ	17	17	الجحيم	23	111.68	**0.72**	(of)-hellfire	6
أَلِيمٌ	38	32	أليم	66	354.65	**0.6912**	(is)-painful	6
أَلِيمٍ	14	14						
عَظِيمٌ	31	31	عظيم	71	146.37	**0.6849**	great	1
عَظِيمٍ	18	18						
غَفُورٌ	51	4	غفور	72	416.21	**0.0676**	oft-forgiving	1
الظَّالِمُونَ	24	16	الظالمون	24	133.1	**0.6538**	the-wrongdoers	6
الْمُجْرِمِينَ	19	12	المجرمين	19	104.72	**0.6190**	[the]-criminals	6
الْمُتَّقِينَ	23	14	المتقين	23	116.18	**0.60**	the-pious	6

Morph. Variant	Raw Count: Variant	Raw Count: Variant with Madd	Unmarked Form	Total Count: Unmarked Form	LL	LaPlace Point Estimate	Interlinear	Quranic Theme
حَكِيمٌ	35	31	حكيم	55	168.71	**0.5789**	full-of-wisdom	1
حَكِيمٍ	4	1						
الْكَافِرِينَ	52	30	الكافرين	52	276.75	**0.5741**	(to)-the-disbelievers	6
تَعْبُدُونَ	20	11	تعبدون	20	118.74	**0.5455**	worship	
عَلِيمٌ	100	56	عليم	128	705.5	**0.4846**	(the)-all-knower	1
عَلِيمٍ	6	6						
بَصِيرٌ	27	20	بصير	42	167.35	**0.4773**	all-seer	1
النَّارِ	70	36	النار	102	156.73	**0.4615**	(in)-the-fire	6
النَّارُ	21	10						
النَّارَ	11	1						
خَبِيرٌ	23	13	خبير	37	181.98	**0.4103**	(is)-all-aware	1
خَبِيرٍ	2	2						
الْأَنْهَارُ	40	11	الأنهار	43	213.77	**0.2889**	the-rivers	6
الْأَنْهَارَ	3	1						
شَهِيدٌ	11	8	شهيد	32	140.03	**0.2647**	(a)-witness	5

Morph. Variant	Raw Count: Variant	Raw Count: Variant with Madd	Unmarked Form	Total Count: Unmarked Form	LL	LaPlace Point Estimate	Interlinear	Quranic Theme
الْعَذَابِ	35	9	العذاب	85	370.77	**0.2184**	the-punishment	6
الْعَذَابُ	22	5						
الْعَذَابَ	28	4						
إِسْرَائِيلَ	41	8	إسرائيل	42	153.88	**0.2045**	children-of-Israel	3
الشَّيْطَانِ	19	7	الشيطان	63	174.67	**0.1692**	Satan	7
الشَّيْطَانَ	10	2						
الشَّيْطَانُ	34	1						
بِالْبَيِّنَاتِ	24	3	بالبينات	24	142.93	**0.1538**	with-the-clear-signs	4
الْآيَاتِ	28	3	الآيات	31	144.49	**0.1212**	the-signs	4
ضَلَالٍ	27	3	ضلال	33	157.82	**0.1143**	error	7
السَّمَاءِ	81	9	السماء	109	215.41	**0.0991**	the-sky	4
السَّمَاءُ	15	1						
الْكِتَابَ	78	5	الكتاب	163	234.92	**0.0667**	the-book	5
الْكِتَابِ	77	5						
لَآيَاتٍ	29	1	لآيات	29	179.31	**0.0645**	surely-signs	4

Morph. Variant	Raw Count: Variant	Raw Count: Variant with Madd	Unmarked Form	Total Count: Unmarked Form	LL	LaPlace Point Estimate	Interlinear	Quranic Theme
سَبِيل	6	5	سبيل	116	296.15	**0.0593**	way/path	5
سَبِيلٍ	7	1						
خَيْر	78	8	خير	153	185.58	**0.0581**	better	U
الصَّالِحَات	59	2	الصالحات	61	335.46	**0.0476**	good-deeds	2
عَذَابٌ	13	1	عذاب	188	969.54	**0.0158**	(a)-punishment	6
عَذَابِ	14	1						
يَوْم	28	2	يوم	233	189.99	**0.0128**	(a)-day	6
السَّمَاوَاتِ	175	1	السماوات	182	1094.65	**0.0109**	the-heavens	4

Appendix III: Comparative rankings of keyness and prominence

La Place Ranking	Unmarked Form	LL Rank	Literal English Translation	Quranic Theme
1	العالمين	17	the-worlds	Nature
2	صادقين	22	truthful (m.pl)	Eschatology
3	خالدون	36	abiding-forever (m.pl)	Eschatology
4	مؤمنين	18	believers/believing	Eschatology
5	المحسنين	38	the-doers-of-good	Eschatology
6	الحكيم	31	the-(all)-wise	God
7	مبين	4	clear	Prophethood/Revelation
8	المرسلين	43	the-messengers	Prophethood/Revelation
9	قدير	16	(all)-powerful	God
10	العليم	25	the-(all)-knowing	God
11	مستقيم	34	straight	Prophethood/Revelation
12	**الظالمين**	**10**	**the-wrongdoers**	**Eschatology**
13	رحيم	5	merciful	God
14	الجحيم	41	the-hellfire	Eschatology
15	أليم	8	painful	Eschatology
16	عظيم	32	great	God
17	الظالمون	39	the-wrongdoers	Eschatology
18	المجرمين	42	the-criminals	Eschatology
19	المتقين	40	the-god-fearing	Eschatology
20	حكيم	26	wise	God
21	الكافرين	12	the-disbelievers	Eschatology
22	عليم	3	(all)-knowing	God
23	**بصير**	**27**	**(all)-seeing**	**God**
24	**النار**	**29**	**the-fire**	**Eschatology**
25	**خبير**	**21**	**(all)-aware**	**God**
26	الأنهار	15	the-rivers	Eschatology
27	شهيد	37	(a)-witness	Prophethood/Revelation
28	العذاب	7	the-punishment	Eschatology

La Place Ranking	Unmarked Form	LL Rank	Literal English Translation	Quranic Theme
29	**إسرائيل**	**30**	**Israel**	**Man in Society**
30	الشيطان	24	the-Devil/Satan	Satan/Evil
31	**بالبينات**	**35**	**with-the-clear-signs**	**Nature**
32	**الآيات**	**33**	**the-signs**	**Nature**
33	**ضلال**	**28**	**error**	**Satan/Evil**
34	السماء	14	the-sky	Nature
35	غفور	6	forgiving	God
36	الكتاب	13	the-book	Prophethood/Revelation
37	لآيات	23	for-signs	Nature
38	سبيل	11	way/path	Prophethood/Revelation
39	خير	20	good	Eschatology
40	الصالحات	9	the-good-deeds	Man (Individual)
41	عذاب	2	punishment	Eschatology
42	يوم	19	day	Eschatology
43	السماوات	1	the-heavens	Nature

Acknowledgements

This research has emerged from an interdisciplinary project funded by the Engineering and Physical Sciences Research council (EPSRC): Atwell, E., Dickins, J. and Brierley, C. (2013). *Natural Language Processing Working Together with Arabic and Islamic Studies* [EP/K015206/1].