## The Quran: The Lexical Profile of the Suras

The aim with the present series, *The Quran: The Lexical Profile of the Suras*, is to present key data related to the lexicon of the suras of the Quran, in terms of Key Word distribution and lexical associations within each sura.

The digital text used for this purpose is the *Uthmani* text of the Tanzil Quran Text (for attribution see below). This text is widely used, and we have conducted some comparison to pre-digital age printed editions of the Quran. All vocalized Arabic text is quoted unaltered in any shape or form from the Tanzil text. Unvocalized Arabic text and transcriptions are my own.

In this volume, covering suras: 12-15, the key data relevant to each ***Sura*** and each ***Key Word*** (KW), here *adjectives, nouns, proper nouns and verbs*, within each sura are presented according to the following:

The Sura: *Number of ayas, Number of Words (and basmala where appropirate), Weight of Sura in the Quran, Number of Key Words, Number of Unique Key Words, Critical Value of Pearson's r for each Sura, Number of Words in each Aya, Distribution of Key Words Unique to the Sura, Key Words Unique to the Sura, Key Word Frequencies*

The Keys Words of the Sura: *Number of Attestations, Weight in Sura, Rank (relative to weight), Distribution and Weighted Distribution by Aya*, *Attested Forms*, *Correlations* and *Collocations all by Sura*.

The Key Words are always referenced by their ***lemma*** and are sorted alphabetically according to Arabic and UNI-

CODE order with minor adjustments for consistency. In lemmatizing the words, no attention has been given to the semantics of each word. Only on rare occasion have similar forms of words or proper nouns been separated in order to avoid confusion.

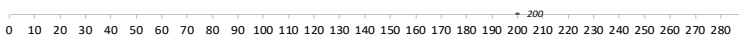
In assigning each word a *lemma*, Classical dictionaries and Quran commentaries, as well as modern Quran dictionaries have been consulted. Deciding on these is not always obvious, since classical dictionaries and commentaries sometimes either disagree or present divergent variant readings or root and lemma attributions.

In the present series, the following definitions and computations are applied and the results presented:

• Weighted Distribution:
nr. of attestations of the KW in each aya
=> weighted distribution of the KW in each aya

Weighted Distribution (WD) is computed as follows:
WD = √w *(in order to accommodate extreme variation)*
w = nr. of attestations of KW in aya /
total nr. of words in aya

The ***Weighted Distribution*** of a certain Key Word represents therefore the percentage that each Key Word occupies in an aya relative to the total nr. of words in that aya. The Weighted Distribution of Key Words across the ***ayas*** of a sura is represented graphically. The following graph plots the weighted value of each Key Word in the ayas of each sura in which it is attested. The dot followed by a number on the x-axis represents the number of ayas in each sura.



• Weight of Sura in the Quran

The ***weight of a sura*** in the Quran is defined as the cumulative value of the weight of every KW in each aya of the respective sura. The resultant sum is divided by the total number of ayas in the Quran (6236), thus providing a percentage of the total weight of the whole Quran.

• Number of Key Words and Unique Key Words

The number of words in a sura represents all the words in a sura, irrespective of word class. The number of ***KWs*** represents the number of words which are *adjectives*, *nouns*, *proper nouns* or *verbs*. The number of ***Unique KWs*** is the number of different KWs irrespective of frequency.

• Critical Value of Pearson's *r* for each Sura

***r.Critical*** represents the critical values of Pearson's *r* for each sura, as described below under Correlations and the table.

• Words per Aya

The number of words per aya for each sura are represented graphically with the grey lines, while the black line represents a *Moving Average* with a *Period* of 5.

• Key Words Unique to a Sura

KWs that are attested in only one sura are listed under the respective sura, and their weighted distribution is presented graphically. These are then tested for correlation (Pearson, see below Correlations and the table) to the length of the ayas. If *r* is greater than r.Critical, then the correlation is statistically significant.

• Weight of Key Word in a Sura:
The ***weight of a KW*** in a sura is defined as the cumulative value of the weight of that KW in each aya of the respective sura in which it is attested. The resultant sum is divided by the total number of ayas in the sura, thus providing a percentage of the total weight of the whole sura. ***Rank*** is the assigned to each KW based on its weight relative to the other KWs in the sura.

• Attested Forms of the Key Words:
The attested forms and frequency of each Key Word in the respective suras are listed alphabetically.

• Correlations:
Correlations are relative distribution patterns of Key Words within each ***sura***. Correlations are calculated with Pearson's Correlation Coefficient *r* for each Key Word attested in a sura relative to every other attested Key Word in the same sura. In the present work, correlations are based on the ***Weighted Distribution*** (as defined above) of each Key Word in every ***aya*** in which it is attested within a sura. *r* ranges from -1 (100% inversely correlated) to 1 (100% correlated). The (statistically significant) critical value for $n$=nr. of ayas in each sura is $|r| \geq$r.Critical (see table below), where $\alpha$=0.05 and $p \leq 0.05$. The null and alternate hypotheses (H0 & Ha) are as follows:

H0: the distribution of KW1 is statistically different from KW2
Ha: the distribution of KW1 is statistically similar to KW2

Key Words that have a relative correlation coefficient that is less than the critical value |r|<r.Critical (H0; not statistically significant) are NOT included under the main lemma entry for each Key Word. The list of statistically significant

correlated Key Words in a sura is sorted first by the degree of correlation *r*, highest to lowest, and then alphabetically according to Arabic and UNICODE order.

• Collocation Frequencies:
Collocation is defined as a Key Word included in a cluster with other Key Words forming the center of that cluster with **six** co-occurring Key Words (*1°, 2° and 3° of proximity*), the first three to the left and the first three to the right, where available. The co-occurring Key Words in a ***sura*** are listed by lemma and are sorted first by the number of co-occurrences (Collocation Frequency), highest to lowest, and then alphabetically according to Arabic and UNICODE order.

It is our hope and aim that this series contributes to Computational Linguistics and Digital Humanities in general, and Computational Linguistics research on the Quran in particular.

Elie Wardini,
Stockholm, January 2024