OSMAN – A Novel Arabic Readability Metric

Mahmoud El-Haj and Paul Rayson
School of Computing and Communications, Lancaster University, UK
{m.el-haj, p.rayson}@lancaster.ac.uk

Abstract

We present OSMAN (Open Source Metric for Measuring Arabic Narratives) - a novel open source Arabic readability metric and tool. It allows researchers to calculate readability for Arabic text with and without diacritics. OSMAN is a modified version of the conventional readability formulas such as Flesch and Fog. In our work we introduce a novel approach towards counting short, long and stress syllables in Arabic which is essential for judging readability of Arabic narratives. We also introduce an additional factor called "Faseeh" which considers aspects of script usually dropped in informal Arabic writing. To evaluate our methods we used Spearman's correlation metric to compare text readability for 73,000 parallel sentences from English and Arabic UN documents. The Arabic sentences were written with the absence of diacritics and in order to count the number of syllables we added the diacritics in using an open source tool called Mishkal. The results show that OSMAN readability formula correlates well with the English ones making it a useful tool for researchers and educators working with Arabic text.

Keywords: Arabic, readability, diacritics, OSMAN, flesch, fog, parallel, corpus, NLP, corpus linguistics

1. Introduction and background

Readability measures the ease with which a written text can be understood by readers from different educational levels. Early research by Kitson (1921) showed that sentence and word length were best indicators of a text being easy to read. Other factors include number of characters, the count of hard and complex words as well as the number of syllables in each word. Since then several popular readability metrics have been introduced such as Dale-Chall (Dale and Chall, 1948) Fog (Gunning, 1968) and Flesch (Kincaid et al., 1975).

Readability formulas predict reading difficulty associated with text. This helps educators in selecting the appropriate text for their audience. The majority of the current readability formulas work with text written in English (Gunning, 1968; McLaughlin, 1969; Kincaid et al., 1975) by counting the number of syllables in the words. Applying those formulas to other languages is more problematic, especially when syllables are not easy to count. For other languages such as Arabic the only current solution is to use language independent metrics such as Laesbarheds-Index (Lix) (Björnsson, 1968) and Automated Readability Index (ARI) (Smith et al., 1967). These metrics are restricted to other factors such as number of words and sentences in addition to the number of long words[1] but do not consider any language specific factors. Arabic is highly inflectional and derivational, therefore word length and number of sentences cannot be used as the only indicators of text readability (Al-Ajlan et al., 2008).

Previous work on Arabic readability and Natural Language Processing has suggested the removal of diacritics to simplify the language processing but this results in missing vital information about the syllables leading to pronunciation ambiguity (Habash et al., 2007). Pronunciation in the absence of diacritics could be challenging even for native Arabic speakers and in particular problematic in proper names and other words associated with locations. For example, the Arabic name "حسن" is mostly associated with the masculine name "Hasan" even though the name could be feminine by changing the pronunciation to "Husn". This can be easily inferred having the diacritics added to the word as in "حَسَن" and "حُسْن" respectively. Similarly the common confusion between "Amman" the capital of Jordan and the country "Oman", both in Arabic have the exact spelling but can be identified with diacritics added in as in "عَمَّان" [Amman] and "عُمَان" [Oman].

The remainder of the paper is organised as follows. We review related work in section 2. and summarise the main features of Arabic script in section 3. which further motivates the requirement for a new readability measure for Arabic. In section 4., the multilingual United Nations corpus is described, from which we selected our data for the experiment. Our novel Arabic readability metric and a comparison with factors used in other readability metrics is presented in section 5. The evaluation results and discussion appear in section 6. Section 7. concludes.

2. Related Work

There has been a growing interest in Arabic language processing and translation, but despite this increasing interest only a few researchers have tackled the problem of finding a proper Arabic readability index.

Al-Rashidi (2005) experiments on the readability of 4th grade school books in Kuwait showed low readability scores falling in what they called "the depression level of readability". They also found that the texts in the Arabic language textbook are not graded according to level. There was no difference between female and male students in terms of the readability level with all the participants agreeing on the difficulty of the textbook material. The experiments found that Arabic textbooks taught for 4th grade students were unsuitable for their age showing the need to adopt readability as an indicator when preparing textbooks. The study was conducted with 632 students from 26 classes in a number of schools from different areas in Kuwait by asking the students to read parts of the textbooks and manually score their readability level.

[1] Words with more than five letters.

Earlier research by Al-Heeti (1984) examined readability by only looking into the average word length. Al-Heeti's[2] formula calculates as:

$$Heeti = (AWL \times 4.414) - 13.468$$

Although appealingly simple, the formula does not work as a good indicator of Arabic text readability, especially given that Arabic is a highly inflectional and derivational language and word length by itself does not reflect difficulty.

Schwarm and Ostendorf (2005) used a statistical language model to train SVM classifiers in order to classify text for different grade levels building a classifier with a variable precision of 38% to 75% that is dependent on the grade level. Similarly, Ajlan et. al (2008) research on Arabic readability used word length along with numbers of characters and sentences and compared their method to other language independent measures such as ARI and LIX.[3] Ajlan presented ARABILITY, an Arabic Readability tool, using machine learning. In their work they stripped out diacritics and counted the number of syllables by estimating the number of vowels in the documents divided by the total number of words. Their estimation does not work as expected here since counting the number of syllables in a language like Arabic is dependent on the presence of diacritics.

Al-Tamimi et. al (2013) introduced AARI, an automatic readability index for Arabic. In their work they extracted seven features which they used to calculate readability, including the number of characters, words, sentences and difficult words.[4] Since AARI does not include the use of any Arabic specific characteristics this makes the method close in function to the original ARI formula, where the number of characters and average word count plays a crucial role in the output score.

In previous work (El-Haj et al., 2014b) we tested whether language independent readability scores correctly preserve the variation in style and complexity in the original text when translating into other languages at various document levels including chapters and parts. We hypothesised that the readability scores for each block of text in the original and translated versions should be similarly ranked if the translation quality is good. In this paper we adopt the same hypothesis in determining whether our Arabic readability metrics correlate well with the English metrics. We base our hypothesis on the fact that the Arabic translation of the UN English resolutions is of high quality being translated by UN language professionals.[5] We hypothesise the Arabic readability scores should correlate well with the English ones when no language dependent factors are being considered and lower otherwise.

In our paper, we use additional features to improve on the previous work on Arabic readability by counting the number of Arabic syllables in the text. In contrast to Al-Tamimi et. al (2013) we include diacritics in order to correctly count the number of syllables in a word. We used Mishkal[6], an open source tool to add diacritics to Arabic text with an accuracy over 85% (Azmi and Almajed, 2013; Bahanshal and Al-Khalifa, 2012), and evaluated our new metric by correlating the readability of high quality translations in a parallel English-Arabic corpus.

3. Arabic Script

To measure readability of Arabic text one needs to understand the characteristics of the Arabic script. Arabic vocabulary is very rich with a high frequency of heteronyms.[7] For example the word “ذهب” could be used to mean both “gold” and “gone” depending on the context it was used in. Ambiguity to this level can be unraveled through the use of “Tashkil” [forming], which is the process of adding diacritics to written text. Diacritics in Arabic are used as a phonetic guide, especially for non-native speakers. Adding diacritics to the previous example will result in “ذَهَبْ” [Dhahab] and “ذَهَبَ” [Dhahaba] to mean “gold” and “gone” respectively.

Diacritics were formed as a method to transcribe correct reading and they are directly associated with syllables. Counting syllables in Arabic is not a straightforward process as it is not affiliated with certain vowels as in English (e.g. ‘a’, ‘e’, ‘i’, ‘o’ and ‘u’). There are eight diacritics in Arabic with only four associated with vowels: “damma”, “fatHa”, “kasrah” and “shadda”. The sounds of the first three diacritics can be doubled resulting in a ‘aN’ sound at the end of the word (e.g. “أَهْلًا وَ سَهْلًا”, [AhlaN wa sahlaN], “welcome”]), in Arabic this is called “Tanween”. Arabic syllables are divided into two types, long and short. Short syllables are simply a single consonant followed by a single short vowel (e.g. “كَتَبَ” [ka-ta-ba], “he wrote”). A long syllable usually is a consonant plus a long vowel (e.g. “كِتَاب” [ki-taab], “book”) the example shows a short syllable followed by a long one. Stress syllables are those considered as double letters, indicating a double consonants with no vowel in between (e.g. “شَدَّدَ, [shaDDaDa], “he stressed”). In our work we count short, long and stressed syllables.

Diacritics are, most of the time, omitted from the Arabic text, which makes it hard to infer the word's meaning and therefore, it requires complex morphological rules to tokenise and parse the text (El-Haj et al., 2014a). In order to correctly calculate Arabic readability we need to count the number of syllables. An Arabic word could contain more than three syllables and still be considered a non–complex word (Al-Ajlan and Al-Khalifa, 2010). To illustrate this we analysed two million Arabic words and found the average word length and syllable count to be five and four respectively. When counting the syllables we only considered the 4 diacritics in Figure 1 in addition to the Tanween. We did not consider the diacritics when calculating word length.

[2] AWL: Average Word Length.

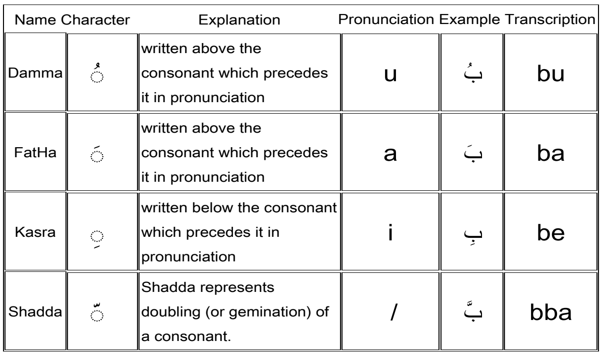
[3] LIX originally worked with Swedish text.

[4] Words with more than six letters.

[5] http://www.un.org/en/sections/about-un/official-languages/index.html

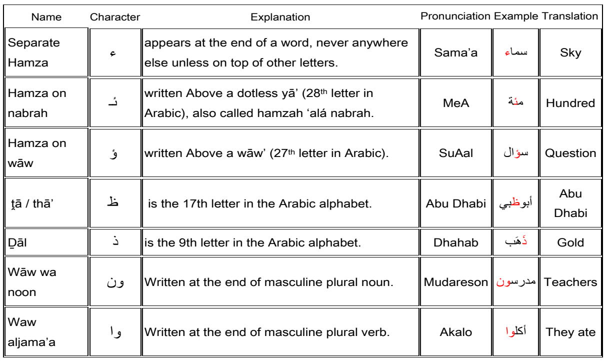
[6] http://sourceforge.net/projects/mishkal/

[7] Identical words that have different pronunciation and meaning



Name	Character	Explanation	Pronunciation	Example	Transcription
Damma	◌ُ	written above the consonant which precedes it in pronunciation	u	بُ	bu
FatHa	◌َ	written above the consonant which precedes it in pronunciation	a	بَ	ba
Kasra	◌ِ	written below the consonant which precedes it in pronunciation	i	بِ	be
Shadda	◌ّ	Shadda represents doubling (or gemination) of a consonant.	/	بَّ	bba

Figure 1: Vowel diacritics



Name	Character	Explanation	Pronunciation	Example	Translation
Separate Hamza	ء	appears at the end of a word, never anywhere else unless on top of other letters.	Sama’a	سماء	Sky
Hamza on nabrah	ئـ	written Above a dotless yā’ (28th letter in Arabic), also called hamzah ‘alá nabrah.	MeA	مئة	Hundred
Hamza on wāw	ؤ	written Above a wāw’ (27th letter in Arabic).	SuAal	سؤال	Question
ṭā / thā’	ظ	is the 17th letter in the Arabic alphabet.	Abu Dhabi	أبوظبي	Abu Dhabi
Ḏāl	ذ	is the 9th letter in the Arabic alphabet.	Dhahab	ذَهَب	Gold
Wāw wa noon	ون	Written at the end of masculine plural noun.	Mudareson	مدرسون	Teachers
Waw aljama’a	وا	Written at the end of masculine plural verb.	Akalo	أكلوا	They ate

Figure 2: Faseeh letters

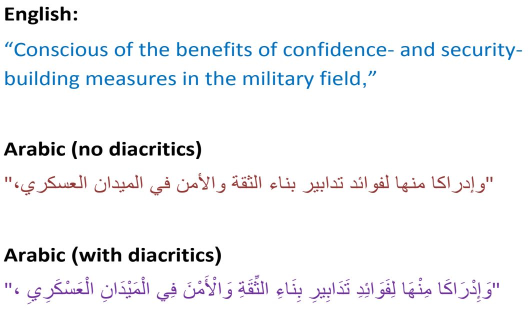
The experiment by Zayed (2006) examined common morphological, grammatical and spelling errors when writing in Arabic. The experiment included 1,200 school students on grade levels 7 to 9 of the Jordanian school system. The experiments measured the students’ proficiency in Arabic writing. Based on their results we selected a number of letters and suffixes and used them to judge readability. The selected letters and affixes shown in Figure 2 are usually misspelled affecting pronunciation and therefore text readability. Most of these spelling errors relate to the influence of the different Arabic dialects. For example, some native Arabic speakers could write in Arabic eliminate all types of Hamza[8] resulting in ambiguity in some contexts.

Originally the Arabic alphabet consisted of twenty-two letters inherited from the Phoenician alphabet, six letters less than the alphabet used nowadays. Letters ‘ذ’, and ‘ظ’ are part of the six letters added to the originally twenty-two letter alphabet. The remaining four letters are ‘ث’,‘خ’, ‘غ’ and ‘ض’. In some cases ‘ذ’, and ‘ظ’ are mistakenly replaced by ‘ز’, and ‘ض’ with the latter being more common. Arabic scholar Al-Khawlani[9] mentioned ninety-three words in Arabic that contain the letter ‘ظ’ with 32 of them are still common nowadays.

Words ending with ‘وا’ or ‘ون’ refer to masculine verbs and nouns, one common error is usually forgetting to add alif ‘ا’ at the end of a word ending with ‘وا’ or using ‘ين’ instead of ‘ون’. Throughout the paper we refer to those misspelled letters as “Faseeh”. Misspelling those letters could result in prosaic Arabic “Rakeek” – weak pronunciation – and therefore affect text readability.

4. Dataset and method

In our experiment, we used 73,000 parallel English and Arabic paragraphs from the United Nations (UN) corpus[10] – a collection of resolutions of the General Assembly from Volume I of GA regular sessions 55-62 (Rafalovitch and Dale, 2009). The Arabic text by the UN has been written with the absence of diacritics. We used Mishkal[11] to add diacritics to the Arabic text. Each language has around 3 million words from more than 2,000 documents with each document containing 36 paragraphs on average.[12] The dataset was originally in TMX[13] file format. We parsed the files and extracted the parallel information and added identifiers to make distinguishing the files a much easier process. The identifiers include document (docID), sentence (sentID) and language (lang), which we used in the publicly available version.[14] Each file follows the following naming patters [docID_sent.lang] (e.g. 1_3.ar).[15] Figure 3 shows a sample of the parallel dataset. Parallel paragraphs share the same docID and sentID across languages.



English:

“Conscious of the benefits of confidence- and security-building measures in the military field,”

Arabic (no diacritics)

"وإدراكا منها لفوائد تدابير بناء الثقة والأمن في الميدان العسكري،"

Arabic (with diacritics)

"وَإِدْرَاكًا مِنْهَا لِفَوَائِدِ تَدَابِيرِ بِنَاءِ الثِّقَةِ وَالْأَمْنَ فِي الْمَيْدَانِ الْعَسْكَرِيِ ،"

Figure 3: Dataset sample

5. Readability Metrics and Methodology

In order to evaluate our work we used several readability metrics in addition to our new Arabic readability score – OSMAN. The readability metrics used in our experiments

[8] Hamza is a letter in the Arabic alphabet, representing the glottal stop [ʔ]. Hamza is not one of the 28 Arabic alphabet.

[9] Abi Al-Hasan Ali bin Muhammed Alkhawlani (also know as Alhaddad Almuhdawi).

[10] http://www.uncorpora.org/

[11] http://sourceforge.net/projects/mishkal/

[12] The resulting dataset as parallel documents and sentences can be downloaded directly from http://drelhaj.github.io/OsmanReadability/.

[13] Translation Memory eXchange

[14] The corpus is freely available for research purposes as indicated by original authors on http://www.uncorpora.org/

[15] In the distributed documents ‘ar’, ‘art’, and ‘en’ refer to Arabic, Arabic with Tashkeel (diacritics) and English respectively.

are Laesbarheds-Index (LIX) (Björnsson, 1968), Automated Readability Index (ARI) (Smith et al., 1967), Flesch Reading Ease and Flesch-Kincaid Grade Level (Kincaid et al., 1975), and Gunning Fog (Gunning, 1968). We modified the formulas to work properly for Arabic by counting words using the Stanford Arabic tokenizer.[16] Here, we do not report results from Flesch-grades (Kincaid) or Fog as these formulas are close to Flesch, we still include details on the formulas as they are presented in the open source code.

The readability metrics formulas are as follows:

$$LIX = \frac{A}{B} + \frac{C \times 100}{A}$$

$$Fog = 0.4 \times \left[\left(\frac{A}{B} \right) + 100 \left(\frac{G}{A} \right) \right]$$

$$ARI = 4.71 \times \frac{E}{A} + 0.50 \times \frac{A}{B} - 21.43$$

$$Flesch = 206.835 - 1.015 \times \left(\frac{A}{B} \right) - 84.6 \times \left(\frac{D}{A} \right)$$

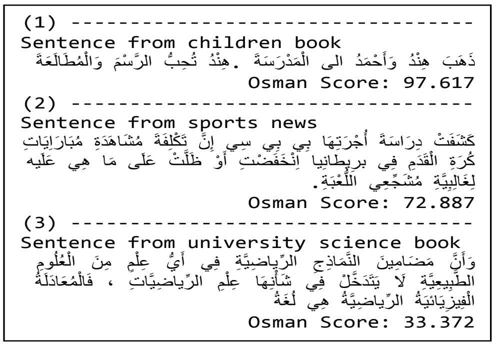
$$Flesch_g = 0.39 \times \left(\frac{A}{B} \right) + 11.8 \times \left(\frac{D}{A} \right) - 15.59$$

Starting with the UN parallel corpus (see Section 4.). We stripped out any empty sentences and sentences with less than three characters. We then replicated the Arabic sentences by adding in diacritics using Mishkal which was trained using the Tashkeela corpus which consists of more than 6 million diacriticised words from a large variety of Arabic textbooks.[17] We calculated readability for the English text using the conventional readability metrics. For the Arabic text, we calculated readability using the same formulas by adapting them to Arabic using the Arabic Stanford tokenizer to count words. We used the diacritics to count syllables. For both versions of the Arabic text we calculated readability using our OSMAN readability formula below which is an updated version of the Flesch formula.

$$Osman = 200.791 - 1.015 \times \left(\frac{A}{B} \right) - \\ 24.181 \times \left(\frac{C}{A} + \frac{D}{A} + \frac{G}{A} + \frac{H}{A} \right)$$

where 'A' is the total number of words counted using Stanford Arabic word tokenizer[18], 'B' is the total number of sentences counted automatically using common delimiters to split text into sentences[19], 'C' is the number of hard words (words with more than 5 letters – long words); the word length was counted with the absence of diacritics in order to avoid counting the diacritics as letters, 'D' is the number of syllables in a word, 'E' is the total number of characters ignoring digits, 'G' is the number of complex words in Arabic (words with more than four syllables as measured by our analysis), 'H' is the number of "Faseeh" words (complex word with any of the following Arabic letters ('ء', 'ىء', 'وء', 'ذ', 'ظ') or ending with ('وا', 'ون'). As these letters could appear frequently we only consider counting Faseeh in complex words, which indicates a higher level of readability. In the formula we count Faseeh by considering the number of complex words containing at least one Faseeh indicator.

The OSMAN formula has been set up to work with plain and diacriticised Arabic text. We use the same formula for both versions, the absence of syllables and complex words in the non-diacriticised text results in higher OSMAN score, which erroneously indicates the text is easier to read, therefore we recommend using diacriticised text or Mishkal to add diacritics to a text before calculating OSMAN score. In order to match the expected distributions across the English and Arabic readability metrics we started by calculating readability for each of the English and Arabic documents in our sample using Flesch with its original arithmetic factors (206.835 and 84.6) but using additional linguistic factors for the Arabic version of the formula as described above. Figure 4 shows sample sentences from a variety of other texts and their OSMAN score, in order to show the range of scores obtained.



```
(1) ---------------------------------
Sentence from children book
ذَهَبَ هِنْدُ وَأَحْمَدُ الى الْمَدْرَسَةَ .هِنْدُ تُحِبُّ الرَّسْمَ وَالْمُطَالَعَةَ
                     Osman Score: 97.617
(2) ---------------------------------
Sentence from sports news
كَشَفَتْ دِرَاسَةَ أُجْرَتِهَا بِي بِي سِي إِنَّ تَكْلِفَةَ مُشَاهَدَةِ مُبَارَاِيَاتِ
كُرَةِ الْقَدَمِ فِي برِيطانِيا اِنْخَفَضْتِ أَوْ ظَلَّتْ عَلَى مَا هِي عَلَيه
لِغَالِبِيَّةِ مُشَجِّعِي اللُّعْبَةِ.
                     Osman Score: 72.887
(3) ---------------------------------
Sentence from university science book
وَأَنَّ مَضَامِينَ النَّمَاذِجِ الرِّيَاضِيَّةِ فِي أَيِّ عِلْمٍ مِنَ الْعُلُومِ
الطَّبِيعِيَّةِ لَا يَتَدَخَّلْ فِي شَأْنِهَا عِلْمِ الرِّيَاضِيَّاتِ ، فَالْمُعَادَلَةُ
الْفِيزِيَائِيَةُ الرِّيَاضِيَّةُ هِي لُغَةُ
                     Osman Score: 33.372
```

Figure 4: OSMAN Score Examples

Taking the UN English documents and Flesch results as gold-standards our aim was to make the distribution of the Arabic texts' (with diacritics) readability scores match the distribution of the English documents' scores. Flesch scores indicate UN text to be best understood by university graduates. We tuned the factors in the OSMAN formula to generates scores between 0 and 100 but as with Flesch the scores can be negative or greater than 100. Using the R statistical package we calculated the density distribution of the resulting readability metrics. Figure 5 shows the distribution of the readability metrics before tuning the arithmetic factors. To reduce the gap between the distributions we calculated the median for the readability scores for each set of documents then updated the Arabic formula by reducing the arithmetic factor by the difference between the Arabic (with and without diacritics) and the English scores median. Recalculating OSMAN readability with the new arithmetic factors resulted in a closer distribution as shown in Figure 6, and this tuned version is the one presented above. As expected, calculating OSMAN for non-diacriticised text results in higher scores due to syllables and complex words counts being zero. We can therefore conclude that including diacritics helps to reflect the actual ease of reading of Arabic text.

6. Evaluation Results and Discussion

To evaluate our new metric further, we compared the readability scores for the UN corpus Arabic paragraphs with and without diacritics to the parallel English paragraphs. We calculated readability metrics for the Arabic and English text using the readability measures mentioned earlier aside from OSMAN score for the English text.

[16] http://nlp.stanford.edu/software/tokenizer.shtml

[17] http://en.sourceforge.jp/projects/sfnet_tashkeela/

[18] http://nlp.stanford.edu/software/parser-arabic-faq.shtml

[19] We also provide the user with an option to choose Stanford Sentences Splitter http://nlp.stanford.edu/software/tokenizer.shtml

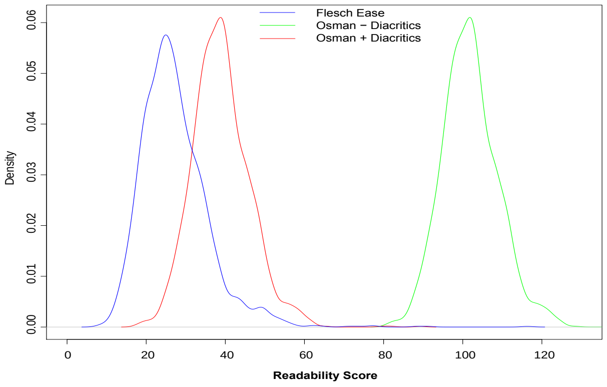


Figure 5: Density Distribution before Tuning

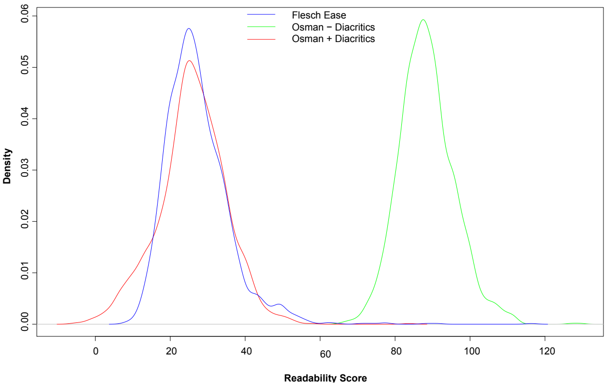


Figure 6: Density Distribution after Tuning

To test the similarity of the readability scores between English and Arabic and between diacriticised and non-diacriticised Arabic text we used Spearman's Rank Correlation Coefficient. Our assumption is that parallel translations should preserve the readability scores, or at least the same rank ordering of scores for the paragraphs. For this purpose we used language independent readability methods. The results confirm our expectations with the language-independent metrics (ARI and LIX) appearing to be fairly consistent between the English and Arabic translations (scores of 0.764 and 0.548 respectively). Table 1 shows the correlation metric scores.

Metrics	Spearman's Score
OSMAN: ART vs AR	0.035
OSMAN ART vs Flesch EN	0.329
ARI: AR vs EN	0.764
LIX: AR vs EN	0.548
Flesch: AR vs EN	0.439
Flesch: ART vs EN	0.221

Table 1: Correlation Scores
(AR: Arabic, ART: Arabic with Tashkeel (diacritics), EN: English.)

The correlation scores drop when linguistic factors are part of the formula. This can be observed by considering the Flesch scores in particular where there is lower correlation between Arabic and English (0.439) and for English compared with the Arabic diacriticised version (0.221). The Flesch formula applied to the Arabic text will not produce accurate results since the syllables count will always be zero, thus giving a higher Flesch score and therefore erroneously indicating the text is easy to read, which is not the case with the UN data.

Spearman's scores showed very low correlation between OSMAN scores of the two Arabic versions (0.035) which indicates the importance of the presence of diacritics, which plays a vital role in determining the text ease of reading. On the other hand the OSMAN scores (0.329) for the diacriticised Arabic showed a higher positive correlation with the English Flesch scores.

This was noticed when measuring the Flesch score difference between the two Arabic texts. We found the diacriticised text to score lower (harder to read) when there are more syllables. This is not surprising as the UN resolution data are of high quality of translation and strict adherence to editorial conventions (Rafalovitch and Dale, 2009).

We measured mean and standard deviation of each metric on each version of the corpus. We would expect that good measures should show the same variability, and for the UN corpus which is fairly homogeneous and that this variability should be quite low. Here again, our OSMAN measure performs consistently on the diacriticised Arabic (mean: 25.89 and stdev: 9.08) with the Flesch on English (mean: 27.26 and stdev: 8.63). The non-diacriticised Arabic mean (88.49) shows how the absence of diacritics (thus syllables and complex words) resulting in inaccurate scores indicating the text to be easy to read in contrast to both Flesch and OSMAN with diacritics. We observe again that the language independent measures (ARI and LIX) are not able to make this distinction.

In terms of a complementary qualitative evaluation, we asked two groups of native Arabic and English speakers (five speakers per language) to read and manually score ten documents on a five point Likert scale. We measured the correlation between the human scores and the scores by OSMAN and Flesch. The scores show high correlation between human and OSMAN scores for the diacriticised Arabic text, which confirms our finding that diacritics are important to determine the text readability. The results also show slightly lower scores for the English text but low correlation between human and OSMAN scores for the non-diacriticised text. Figure 7 shows the scores by the human participants compared to OSMAN for the diacriticised Arabic text confirming with our previous findings on the absence of diacritics.

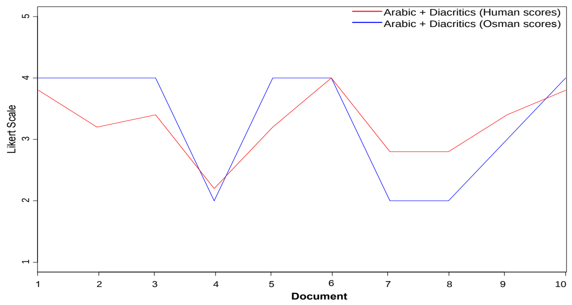


Figure 7: Human vs Osman Readability Scores

7. Conclusion

The paper describes a novel approach for calculating the readability of Arabic text and compares results for text with and without diacritics. The open source Java tool allows users to calculate readability for Arabic text. The tool provides methods to split the text into words and sentences, count syllables, Faseeh letters, hard and complex words in addition to adding and removing diacritics. This makes the tool useful for researchers and educators working with Arabic text. To evaluate the results we used correlation rank and calculated the mean and standard deviation for readability of English and Arabic parallel translations. Spearman's scores showed very low correlation between OSMAN scores of the two

Arabic versions which indicates the importance of the presence of diacritics, which plays a vital role in determining the text ease of reading. On the other hand the OSMAN scores for the diacriticised Arabic showed a higher positive correlation with the English Flesch scores. All the readability metrics mentioned earlier are included within the open source code, they all work with diacriticised and non-diacriticised text but based on the results presented here we recommend adding the diacritics in by using the *addTashkeel()* method. The tool and dataset have been made publicly available for research purposes.[20]

Al-Ajlan, A. and Al-Khalifa, H. (2010). Automatic readability measurements of the arabic text: An exploratory study. *Arabian Journal for Science and Engineering*, 35(2):103 – 124.

Al-Ajlan, A., Al-Khalifa, H., and Al-Salman, A. M. (2008). Towards the development of an automatic readability measurements for arabic language. In *Third International Conference on Digital Information Management ICDIM*, pages 506–511, Nov.

Al-Heeti, K. (1984). Judgment analysis technique applied to readability prediction of arabic reading material.

Al-Rashidi, M. H. (2005). Level of the arabic textbook readability for the basic grade four in the state of kuwait. In *Master Thesis*. The University of Jordan.

Altamimi, A., Jaradat, M., Aljarrah, N., and Ghanem, S. (2013). Aari: Automatic arabic readability index iajit first online publication. In *The International Arab Journal of Information Technology*.

Azmi, A. and Almajed, R. (2013). A survey of automatic arabic diacritization techniques. *Natural Language Engineering (NLE) Journal*, pages 1–19.

Bahanshal, A. and Al-Khalifa, H. (2012). A first approach to the evaluation of arabic diacritization systems. In *Digital Information Management (ICDIM), 2012 Seventh International Conference on*, pages 155–158, Aug.

Björnsson, C. H. (1968). Läsbarhet. In *Stockholm: Liber*.

Dale, E. and Chall, J. (1948). A formula for predicting readability. In *Educational Research Bulletin*, pages 11–28. Taylor Francis, Ltd.

El-Haj, M., Kruschwitz, U., and Fox, C. (2014a). Creating language resources for under-resourced languages: methodologies, and experiments with arabic. *Language Resources and Evaluation*, 49(3):1–32.

El-Haj, M., Rayson, P., and Hall, D. (2014b). Language independent evaluation of translation style and consistency: Comparing human and machine translations of camus' novel "the stranger". In *Text, Speech and Dialogue*, volume 8655 of *Lecture Notes in Computer Science*, pages 116–124. Springer International Publishing.

Gunning, R. (1968). *The Technique of Clear Writing*. McGraw-Hill.

Habash, N., Soudi, A., and Buckwalter, T. (2007). *On Arabic Transliteration*, volume 38 of *Text, Speech and Language Technology*. Springer Netherlands.

Kincaid, J., Fishburne, R., Rogers, R., and Chissom, B. (1975). Derivation of new readability formulas (automated readability index fog count and flesch reading ease formula) for navy enlisted personnel. In *Research Branch Rep*, pages 8–75, Memphis, TN.

Kitson, H. (1921). *The Mind of the Buyer: A Psychology of Selling*. Macmillan.

McLaughlin, H. (1969). SMOG grading - a new readability formula. *Journal of Reading*, pages 639–646, May.

Rafalovitch, A. and Dale, R. (2009). United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. In *Proceedings of the MT Summit XII*, pages 292–299. International Association of Machine Translation, August.

Schwarm, E. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.

Smith, E., Senter, R., and (U.S.), A. F. A. M. R. L. (1967). *Automated Readability Index*. AMRL-TR-66-220. Aerospace Medical Research Laboratories.

Zayed, F. K. (2006). *Common Morphological, Grammatical and Spelling Errors in Arabic*. 2006/6/1747. Yazori Publisher.

[20] http://drelhaj.github.io/OsmanReadability/