Corpus files for the research Paper
Building a Multi-Modal Arabic Corpus (MMAC)

The corpus files contain the following files and folders:

1. **// Original Data File/WORD.txt:** This file is in text format with Windows 1256CP code page. It contains the original words as collected in sequence. It is corrected and has some proof reading. This file can be used for defining word frequency pairs. Be care the file is big and takes much time to open.
2. **// Original Data File After Replacing Alef/ WORD replaced.txt:** This file is the same like the previous one but with the following replacements in letters along the whole document:
    a. Replace final letter (ى) with (ي).
    b. Replace final letter (ة) with (ه).
    c. Replace letters (أ، إ، آ) with (ا).
3. **// Sample Raw Data Collected/:** Contains samples of all types of files that are used in collecting the corpus. The files are on their original formats.
4. **// Naked PAWS Images/:** This folder contains three sub-folders for the main three Arabic font files. These folders are Arabic Transparent Font, Simplified Arabic Font and Traditional Arabic Font. In every folder there are computerized images of the 32,917 unique naked PAWs in .tif image mono file format. Every file contains the image of the PAW and the file name in English transliterated from the Arabic naked PAW in the image. The transliteration code is in the following table.

Transliterated Character	Arabic Character	Transliterated Character	Arabic Character	Transliterated Character	Arabic Character	Transliterated Character	Arabic Character	Transliterated Character	Arabic Character	Transliterated Character	Arabic Character
'	ء	#A	ا	#H	ح	s	س	#E	ع	l	ل	Y	ي
I	آ	B	ب	x	خ	$	ش	g	غ	m	م		
A	أ	P	ة	d	د	#S	ص	-	ـ	n	ن		
@	ؤ	T	ت	~	ذ	#D	ض	f	ف	h	ه		
E	إ	V	ث	r	ر	#T	ط	q	ق	w	و		
}	ئ	J	ج	z	ز	#Z	ظ	k	ك	#Y	ى		

5. **// Data Files/:** This folder contains four sub-folders for the main statistical analysis of the research. These folders are Words, Naked Words, PAWs and Naked Paws. In each folder of the previous folders there are the following files:
    a. **Words.txt:** Includes a list of all the words in the corpus after replacing characters and randomize the file.
    b. **Unique sorted Words.txt:** Includes a list of the unique sorted words of the corpus.
    c. **Word Repetition Counting.txt:** Includes a list of the unique words and the number of repetition of each word.
    d. **Top 50 repeated Words.txt:** Includes a list of the top 50 repeated words in the corpus.
    e. **Word Repeated Once.txt:** Includes a list of the words that repeated once in the corpus.
6. **// Training Dataset Files/:** This folder contains two sub-folders for the typed and scanned documents. Each folder contains 19 documents. You will find a big variety in the shape and quality of the scanned documents. Starting from good quality document to very bad quality once. The documents scanned with resolution 200DPI, 24bit RGB color mode and tiff file format. The typed documents are in the Microsoft Word 2003 file format.
7. **//Frequency List/:** This folder contains the frequency lists of the Words, NWords, PAWs, NPAWs. The lists are sorted alphabetically of words. Each line contains the word with its frequencies. The folder contains four files.
8. **All Statistics.xls:** This file contains all the data sheets and charts that have been used in the corpus data analysis.