

Alphabetical Collation Sequence of Arabic Words With Special Characters in Microsoft Office Software

Manar Almanea, Imam Mohammad Ibn Saud Islamic University, Saudi Arabia

The Asian Conference on Arts & Humanities 2025
Official Conference Proceedings

Abstract

Arabic is a language characterized by a large number of special characters, such as accents and symbols, in its script beyond alphabetic letters. While Arabic adheres to a fixed alphabetical order, the arrangement of words containing these special characters remains controversial. This study investigates the degree of sophistication of the Arabic alphabetical sorting systems operating in Microsoft Office Word and Excel documents, as well as in Python, which employs UTF-8 encoding. A list of 38 Arabic words was used for evaluation purposes. Each group of words in the list shared almost the same consonantal root but with varying characters and diacritics. Extraction and comparison of the sorted outputs from the three programs revealed marked sorting differences in the three sorted lists, with discrepancies as significant as 58% observed across the tested conditions. This is not just a simple technical error—it's a linguistic and cultural oversight with consequences for data integrity and accessibility. Similarities and differences in the orders of the generated lists are then discussed. To solve this problem, this study proposes a linguistically-informed *secondary* alphabetical order for special characters beyond the *primary* order of Arabic letters. The order is based on some linguistic features of the special characters, such as the word's root and the character's phonological salience. Software developers working with Arabic script in digital applications are advised to incorporate the recommendations of this study into their work and to make adjustments to the alphabetical collation algorithms implemented within their programs.

Keywords: Arabic digitization, Arabic special characters, alphabetical sorting, collating system, Microsoft Office

iafor

The International Academic Forum
www.iafor.org

Introduction

The world is witnessing unprecedented advancements in areas that intersect language and computer science, marked by a proliferation of digital linguistic reservoirs and Large Language Models. However, due to specific linguistic features, the written form of some languages is easier to digitize and manipulate electronically than others. Arabic is among the languages whose script conventions exhibit details that complicate digitization. Despite being used by around 440 million people around the world (Motwakel et al., 2023), the fourth used language on the internet (Al-Onazi et al., 2023), and the sixth official language of the United Nations (Al-Onazi et al., 2023), the digitization of Arabic still faces some unresolved issues. An inherent feature of the Arabic script is its extensive repertoire of special characters, roughly 10 symbols, which are intricately integrated with letters to alter their form within words and mainly represent phonological information. They also frequently influence the grammatical and semantic significance of a word. Sometimes, more than one special character is combined with a single letter each conveying a specific meaning e.g. a consonant doubling symbol and a short vowel symbol. The most frequently used special characters in Arabic are presented in Table 1 below, along with their meaning.

Table 1

Examples of Special Characters and Their Linguistic Functions in Arabic

Prolonged vowel (~)	Short vowels (َ ؀ِ ؓ)	indefiniteness (َ ؓ)	Consonant doubling (ّ)	Glottal stop (ء)
------------------------	--------------------------	-------------------------	---------------------------	-----------------------

The symbol (~) over a vowel in Arabic represent a prolonged vowel, i.e., a double vowel that is phonemic in Arabic as the following minimal pair show (مَال - مَال) (money-destination). Similarly, the symbol for consonant doubling (ّ) is phonemic and marks the difference between words differing only in this symbol as (أَلَم - أَلَم) (pain-happened). All short vowels in Arabic—/a/, /i/ and /u/—are not represented through letters in writing but rather through diacritics. These can be within words, or on the last letter representing grammatical case. The symbol (ء) is the glottal stop sign, while these symbols (َ ؓ) represent the indefinite article in Arabic. These symbols represent an integral part of Arabic alphabets.

Languages employ different conventions for treating modified letters and certain letter combinations. For example, in Spanish, the letter ñ is treated as a basic letter (separate character) following n in order, and the digraphs ch and ll were formerly (until 1994) treated as basic letters following c and l, although they are now alphabetized as two-letter combinations (Collation-Charts.ORG, 2024). In Arabic, alphabetical sorting of words with such special characters, i.e., “alphabetical collation sequence” (Shihab, 2006), still lacks a standard in Arabic. Software programs such as operating systems and database management systems order Arabic words with special characters differently. This problem had long been highlighted (since the 1980s) but remains unresolved (Elmaghraby et al., 1989; Mustafa, 1996).

Theoretical Background

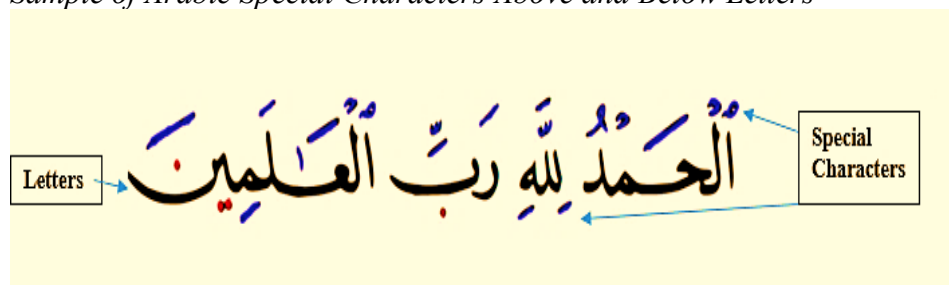
A collating system is the assembly of written information into a standard order (Elmaghraby et al., 1989). Collation algorithms and encoding standards such as Unicode establish a specific order by comparing two character strings and determining their hierarchical

placement. When an order has been defined in this way, a sorting algorithm can be used to put a list of any number of items into that order. The writing systems of certain languages require the decision and announcement about how to order letters accompanied with special characters as in Urdu (Hussain & Karamat, 2003), Spanish, and Croatian (Collation-Charts.ORG, 2024). We believe that such an effort has not been explored in Arabic.

Arabic is a language featuring a large number of special characters (i.e., accents and symbols) beyond alphabets in its script (Ryding, 2005; Ryding, 2014) (see Figure 1).

Figure 1

Sample of Arabic Special Characters Above and Below Letters



These include, but are not exclusive of, short vowel diacritics above or below letters (َ, ِ, ُ), case-marking diacritics at the last letter (َ, ِ, ُ), nunnation at the last letter (ً, ٍ, ٌ) to indicate indefiniteness, the glottal stop symbol (ء) in various positions (أ, إ, ئ, ؤ), and the Prolonged vowels (ـ). Some letters such as Ta' (ة, ت) and Alef (أ, إ) can have various forms at the end of the word¹ (Ryding, 2005; Ryding, 2014; Shihab, 2006). These details create challenges for the digital manipulation of the Arabic script. Although Arabic adheres to a fixed primary alphabetical order of the letters (Figure 2),² the arrangement of words with these special characters (what we refer to in this study as secondary alphabetical order) remains non-unified and lacks a standard. This creates a problem for alphabetical sorting algorithms in Arabic, which require stability as a key feature. Non-standardized sorting algorithms across commonly used software programs (e.g., Word, Excel, Python) create real problems in administrative systems, search engines, lexicons, and education, impacting 440+ million Arabic speakers and users. It is worth mentioning that the Arabic writing system is phonological, where direct correspondence exists between the Arabic written form of words and their pronunciation. Indeed, the written form mirrors exactly how words are pronounced in Arabic. Accordingly, phonological rules can play a role in Arabic sorting systems, as the remainder of this study will show.

¹ The script of the Quran involves much more special characters than regular Arabic writing. They guide the special phonological recitation of verses. They are not all analyzed in this study.

² The letter *Wow* precedes the letter *Ha* in the Western Arabic countries unlike Eastern Arabic countries in which *Ha* preceded the *Wow* (Mustafa, 1996).

Figure 2*Arabic 28 Alphabets Ordered in Columns From Right to Left*

alif	ا	za	ز	qaf	ق
ba	ب	sin	س	kaf	ك
ta	ت	shin	ش	lam	ل
tha	ث	sad	ص	mim	م
jim	ج	dad	ض	nun	ن
ha	ح	ta	ط	ha	ه
kha	خ	dha	ظ	waw	و
dal	د	ain	ع	ya	ي
dhal	ذ	ghain	غ		
ra	ر	fa	ف		

As a root-pattern language, most Arabic dictionaries order entries depending on the trilateral root consonants (Wehr & Cowan, 1996) following the letters' alphabetical order. Another order system was based on the consonants' place of articulation as in *Kitāb al-ʿAyn*, the first Arabic dictionary compiled by Al-Farahidi in the 8th century in which entries were ordered starting from “deepest” pharyngeal letters and ending with bilabial sounds (Al-Makhzūmi & Al-Samirāʾi, 1988).³

Establishing a unified scheme and a collating sequence for electronically sorting words with these special characters is crucial for Arabic users (Shihab, 2006). A collation sequence is vital because it determines the rules for comparing and sorting textual data in databases or large documents. It impacts many computerized operations' search results, data retrieval, consistency, and accuracy. It also ensures a successful migration between various software programs more easily. A unified system is necessary for filing systems, data cataloging, indexing, organization of long name lists, legal documents, medical files, reference books, lexicography, and corpora concordance lists, among other uses. As an example of such a need, educators, teachers, and university professors need a fixed alphabetical order in long lists of students' names, particularly when moving these lists from one platform to another, or when entering grades and or other personal information to ensure accuracy. Following a unified alphabetical order is even considered a force for democratizing access to information (Street, 2020).

Objectives

This study investigates the degree of unification and sophistication of Arabic alphabetical sorting systems operating in two widely used Microsoft Office Software 2021 programs, namely Word (MW henceforth) and Excel (ME henceforth). It examines how these programs deal with words with diacritics and special characters under different conditions during alphabetical sorting. Specifically, the study examines the sorting of:

- Different types of initial Alef (أ - إ - ا)
- Different positions of the glottal stop (ء - و - ئ)
- Short vowel diacritics (َ - ُ - ِ)
- Nunnation (final -n indefiniteness symbol) (ة - ً)

³ An Arabic reference.

- Prolonged vowel (double vowel) (~)
- Gemination (double consonants) (ّ)
- Different final Ta shapes (تـ , تـ)
- Final Ta vs. final Ha (هـ , هـ)
- Different final Alef types (اـ , اـ , اـ)

If a specific sorting scheme is found across these programs, the study discusses and evaluates it accordingly. The generated lists are then compared to a list obtained from Python that utilizes the Universal Unicode (UTF-8) system. Finally, a recommended order of sequencing these special characters in Arabic is proposed. The proposed system will operate based on the linguistic properties of every symbol. This effort aims to provide a standard Arabic collation sequence for special characters.

Research Questions

1. Is there a specific unified alphabetical order followed by MW and ME in sorting words with special characters?
2. Are the lists obtained from MW and ME identical to the order obtained from Python based on UTF-8?
3. What is the recommended system for alphabetically sorting special characters in Arabic?

Methodology

To evaluate how these systems sort words containing the various special characters mentioned above, the researchers, all native Arabic speakers with expertise in linguistic studies, meticulously compiled and revised a list of 38 actual Arabic words (Table 2). These words were clustered into 12 groups sharing the same consonantal trilateral root. Every word addresses the sorting problem of a specific special character (or combination of characters). The selection criteria for words depended mostly on varying the special character under investigation while holding the rest of the word exactly the same. Each group of these words shared almost the same stem or root of a word (almost the same letters) but with varying special characters (examples shown in Table 3). For every group, variation is only in the symbol under investigation.

Table 2*Arabic Words Included in the Study and Their English Meaning*

Arabic word	English meaning	Arabic word	English meaning	Arabic word	English meaning
أبي	My father	سأل	Asked	عَصْرٍ	Age (genitive indefinite)
اثنين	Monday	سَمَا	Surpassed	عَصْرُ	Age (nominative)
إثنين	Two	سَمَاء	Sky	عَصْرًا	Age (accusative indefinite)
اسم	Name	سؤل	Request	عَصَى	disobeyed
إمارة	Emirate/ principality	سئل	Was asked	العِلْم	Knowledge (definite)
أمارَة	token/sign	صديقة	A friend (female)	قَلَم	A pen
أملًا	In hope of	صديقه	His friend (male)	القَلَم	The pen
أملًا	I am in hope of	ضَمَنَ	Included	كَتَاب	A book
الزَّار	Al-Zar (African tradition)	ضَمَنَ	Guaranteed	كَتَبَ	Wrote
زَارَ	Visited	عَصَا	A stick	كُتِبَ	Books
زهرة	Flower	عَصَرَ	Squeezed	كُتِبَ	Was written
الزَّهْرَة	The flower	عَصْرُ	Age (nominative indefinite)	كُتِبَة	Writers
				كُتِبَتْ	She wrote

Table 3*Special Characters Examined in the Present Study*

Special characters	Name in Arabic	Description	Example/ possible minimal pairs	Type of difference indicated
ا - أ - إ	- همزة الوصل وهمزة القطع -على الألف أو تحت الألف	Various forms of the Alef letter (either combined or not combined with the glottal stop symbol)	اثنين - إمارة إمارة	Morphemic (indicating different morphemes)
أ - و - ئ	مواضع الهمزة المختلفة	Glottal stop shape based on surrounding vowels	سأل - سؤل - سؤل	Morphemic (indicating different morphemes)
(َ) (ِ) (ُ)	الحركات على الحروف	Short vowel diacritics above or below the letter	كُتِبَ كُتِبُ كُتِبِ عَصَرَ عَصْرُ	-Morphemic within words (indicating different morphemes) -Allomorphic at the last word

(ء)	التنوين	Nunation- indicator of indefiniteness	عَصْرٌ - عَصْرًا	Nunation is an indefinite morpheme, But the difference between a word with or without it is allomorphic
(~)	المد	Prolonged vowels	أَمَلًا - أَمَلًا	Morphemic
(ة ، ت)	التاء المربوطة والمفتوحة	Letter Ta shapes at the end of the word	كتبة - كتبت	Morphemic
(ا ، ي ، ء)	الألف الممدودة والمقصورة	Letter Alef shapes at the end of the word	عَصَا - عَصَى فنى - فناء	Morphemic
(ة - هـ)	التاء المربوطة والهاء المربوطة	Final Ta vs. final Ha	صديقة - صديقة	Morphemic
(ّ)	الشدة	Gemination	ضَمَنَ - ضَمَّنَ	Morphemic

The list of selected words was inserted into both MW and ME. Additionally, the list was entered in Python, and the program was required to sort the list. The alphabetical order of the list is generated from the three programs (see Table 4).

Results

Table 4

Alphabetically Sorted Lists Generated by MW, ME, and Python

MW		ME		Python (UTF-8)	
Word	Order	Word	Order	Word	Order
أبي	1	أبي	1	أَمَلًا	1
اثنين	2	اثنين	2	أبي	2
إثنين	3	اسم	3	أَمَارَة	3
اسم	4	الْعِلْمُ	4	أَمَلًا	4
إمارة	5	الْقَلَمُ	5	إثنين	5
أَمَارَة	6	الرَّار	6	إمارة	6
أَمَلًا	7	الرَّار	7	اثنين	7
أَمَلًا	8	الرَّهْرَة	8	اسم	8
الرَّار	9	إثنين	9	الرَّار	9
رَار	10	إمارة	10	الرَّهْرَة	10
زهرة	11	أَمَارَة	11	الْعِلْمُ	11
الرَّهْرَة	12	أَمَلًا	12	الْقَلَمُ	12
رَهْرَة	13	أَمَلًا	13	زهرة	13
سأل	14	رَار	14	رَار	14
سَمَا	15	رَار	15	رَهْرَة	15
سَمَاء	16	رَهْرَة	16	سأل	16
سُول	17	زهرة	17	سُول	17

سئل	18	سَمَا	18	سئل	18
صديقة	19	سَمَاء	19	سَمَا	19
صديقه	20	سأل	20	سَمَاء	20
ضَمَّنَ	21	سؤل	21	صديقة	21
ضَمَّنَ	22	سئل	22	صديقه	22
عَصَا	23	صديقة	23	ضَمَّنَ	23
عَصَرَ	24	صديقة	24	ضَمَّنَ	24
عَصْرُ	25	صديقه	25	عَصَا	25
عَصِرَ	26	ضَمَّنَ	26	عَصَرَ	26
عَصْرُ	27	ضَمَّنَ	27	عَصَى	27
عَصْرًا	28	عَصَا	28	عَصْرُ	28
عَصَى	29	عَصَرَ	29	عَصِرَ	29
الْعِلْمُ	30	عَصَى	30	عَصْرًا	30
قَلَمُ	31	عَصْرُ	31	عَصْرُ	31
القَلَمُ	32	عَصِرَ	32	قَلَمُ	32
كِتَاب	33	عَصْرًا	33	كُتِبَ	33
كُتِبَ	34	عَصْرُ	34	كُتِبَ	34
كُتِبَ	35	قَلَمُ	35	كُتِبَتْ	35
كُتِبَ	36	كُتِبَ	36	كُتِبَ	36
كُتِبَ	37	كُتِبَ	37	كُتِبَ	37
كُتِبَتْ	38	كِتَاب	38	كِتَاب	38





Analysis

As indicated in Table 3, variations in the special characters under investigation within words are mostly morphemic in Arabic (with the exception of final letter diacritics and nunnation). This means that the existence of a special character makes a different (independent) word and must be considered by software programs as separate, independent characters with a status equal to a letter. Final letter diacritics and nunnation add only syntactic (case-marking) and semantic (indefinites) information to the word. Variations in them are allomorphic. In other words, variations in these special symbols do not create different words, and they do not require consideration as an independent character. The same applies to words written with short vowel diacritics and those in which the diacritics are not recorded. They can be the same words because short vowel diacritics are optional in Arabic writing. A decision on whether they are the same words or not requires contextual information to solve possible ambiguity. This information is important for the normalization processes needed for some software development.

Table 5 below summarizes the analysis of results obtained by comparing the three alphabetically sorted lists generated from the three programs.

Table 5*Comparison of the Alphabetical Lists Obtained From MW, ME, and Python*

Conditions and special characters	MW	ME	Python	Unification Status
The letter <i>Alef</i> with or without glottal stop ʾ at the word-initial position (<i>Hamzat Alwasl</i> and <i>Hamzat Alqata'</i>) (ا-أ)	ا before أ <i>Hamzat Alqata' First</i>	ا before أ <i>Hamzat Alqata' First</i>	أ before ا <i>Hamzat Alwasl First</i>	Order is different
Words with and without the definite morpheme prefix {-ال}	Does not consider it part of the word; it sorts the words regardless of the existence of the definiteness prefix	Considers it as a part of the word (all definite words are sorted under each other)	Considers it as a part of the word (all definite words are sorted under each other)	Order is different
Glottal stop with the letter <i>Alef</i> at the word-initial position (<i>Hamzat Alwasl</i> and <i>Hamzat Alqata'</i>) when the rest of the word is different	Considers ا - أ one entity	Considers them different, ا is before أ	Considers ا - أ one entity	Order is different
The letter <i>Alef</i> with short vowels: ا - إ	ا before إ	ا before إ	إ before ا	Order is different
Glottal stop at the medial position of the word (medial <i>Hamzah</i>)	The order is: أ ؤ ئ	The order is: أ ؤ ئ	The order is: أ ؤ ئ	Order is the same
Short vowel diacritics (<i>Fatha</i> /a/ <i>Dhammah</i> /u/ <i>Kasrah</i> /i/ <i>Sokon</i>) ه ا إ أ ـ	<i>Fatha</i> /a/ <i>Dhammah</i> /u/ <i>Kasrah</i> /i/ <i>Sokon</i>	<i>Fatha</i> /a/ <i>Dhammah</i> /u/ <i>Kasrah</i> /i/ <i>Sokon</i>	<i>Fatha</i> /a/ <i>Dhammah</i> /u/ <i>Kasrah</i> /i/ <i>Sokon</i>	Order is the same
Words with short vowel diacritics vs. words without short vowel diacritics	Non-diacritized before diacritized	Diacritized before non-diacritized	Non-diacritized before diacritized	Order is different

Varying lengths of words	It sorts words based on the alphabetical order of letters of the words (it compares the first letters, if similar, it compares the second, and so on)	It sorts by comparing vowel diacritics of letters of the words	It sorts by comparing vowel diacritics of letters of the words	Order is different
Indefinite suffix: nunnation				Order is the same
Geminated and non-geminated letters 	Geminated before non-geminated	Non-geminated before geminated	Geminated before non-geminated	Order is different
Prolonged /a/ vowel \bar{a} vs. \acute{a}	\bar{a} before \acute{a}	\bar{a} before \acute{a}	\bar{a} before \acute{a}	Order is different
Final <i>Alef</i> shapes (Mamdodah-Maqsurah) \bar{a} – \acute{a}	\bar{a} before \acute{a}	\bar{a} before \acute{a}	\bar{a} before \acute{a}	Order is the same
Final <i>Ta</i> shapes: (Marbootah-Maftoohah) \bar{t} – \acute{t}	\bar{t} before \acute{t}	\bar{t} before \acute{t}	\bar{t} before \acute{t}	Order is the same
Final <i>Ta</i> vs. final <i>Ha</i> (<i>Ta Marbootah</i> – <i>Ha Marbootah</i>) \bar{t} – \bar{h}	\bar{t} before \bar{h}	\bar{t} before \bar{h}	\bar{t} before \bar{h}	Order is the same

Discussion

A quick glance at the generated ordered lists of words in the three programs in Table 4 shows that none of the lists are identical. Indeed, out of the 14 conditions for sorting, the three programs generated identical order for only 6 conditions comprising 42%. The remaining 58% show varying orders in the three programs (see Table 6). Considering the ranks (1, 2, 3, etc.), no rank (no horizontal line in Table 4) was occupied by the same word in the three programs. This indicates that users must be cautious when migrating long lists between programs.

Table 6*Percentages of Similarities and Differences in the Alphabetically Sorted Lists*

	No.	%
Similarities	6	42%
Differences	8	58%
Total	14	100%

The discussion below is supported by some linguistic details and features of the different special characters. It is divided into two sections: one focusing on similarities between the lists of software programs and the other on their differences.

Similarities

The three programs sorted words with specific special characters similarly in the conditions discussed below.

- The three programs sort **short vowel diacritics** in Arabic (أَ، إ، ا) as follows: *Fatha* /a/, *Dhamah* /u/ then *Kasrah* /i/. This order is similar to the alphabetical order of the corresponding long vowel letters *Alef* (أ) /a:/, *Wow* (و) /u:/ and finally *Ya'* (ي) /i:/. The same order is followed in all programs with different glottal stop shapes within words (أْ- (وْ) –(يْ).
- **Nunnation** that combines the indefinite {-n} with different syntactic case-marking short vowels is ordered the same across the three programs. The order is as follows: nominative (أَ) followed by genitive (إِ) and finally accusative (أِ). Nominative is the default and unmarked case for nouns in Arabic.
- **Final Alef** types are ordered in the three programs as follows: (أ) before (ي). This can be explained linguistically since final (أ) is originally a (و) in the underlying root of the word, while (ي) is originally a (ي) in the underlying root of the word (Elmaghraby et al., 1989). In Arabic alphabetical order (و) precedes (ي).
- **Final Ta** shapes are ordered in all programs as follows: (ت) *Maftooahah*, then (ة) *Marbootah*. This order is justified linguistically since (ت) represents a more stable (invariant) sound as it has no allophones. It is phonetically more salient⁴ than the sound represented by (ة). (ة) has the allophone /h/ (ه) used if the word is followed by a juncture (silence). In addition, the final /h/ is frequently devoiced in speaking in many dialects of Arabic. This results in the possibility of devoicing of (ة) in speech. Furthermore, (ت) is in many cases a part of the underlying trilateral root as in the word بيت “house,” as opposed to (ة), which is often part of inflectional or derivational morphology as the feminine (ة) in صديقة “a female friend” (Halawani, 2017).⁵
- Comparing **final Ta** (ة) with **final Ha** (ه), (ه) always precedes the letter (ة) at the final position, an order aligned with the alphabetical order of the letters (ت) and (ه).

Differences

- Sorting words with initial position *Alef* with and without the glottal stop is variant in the three programs. These are referred to in Arabic as *Hamzat Alwasl* and *Hamzat Alqate'*. The two letters (أْ) and (أ) are not allomorphic. Evidently, the two words (إثنين) and (الثين) only differ in the glottal stop and the first means “Monday” while the other

⁴ Phonological salience refers to prominence or perceptual importance/weight of certain sounds. More salient sounds stand out to speakers and listeners due to some phonetic characteristics such as loudness, duration, stress, syllabic complexity, voicing, etc. (Baroni, 2014; Gussman, 2002).

⁵ An Arabic reference

means “two.” Python orders (ا) before (أ) while MS and ME orders them the opposite way. /أ/ is phonologically more salient (it contains a consonant and a vowel always) than /ا/ which represents only a vowel in continuous speech. Accordingly, it is linguistically justified to begin with *Alef* with glottal stop (أ) than with the one without (ا).

- However, if the two words contain different letters in the rest of the word as in اسم “name” and أبي “father,” MW and Python treat /ا/ and /أ/ as one entity and depend on the following letter to sort the words. Accordingly, أبي precedes اسم based on the alphabetical order of the second letter of the word. In this way, MS and Python yield a fluctuating list that appears to be not systemized. In addition, the two letters are independent as mentioned before. On the contrary, ME sticks to its predisposed sorting of (أ) before (ا). This is a more linguistically justified and systemized choice.
- Sorting initial *Alef* with varying vowel diacritics (أ) and (ا) was problematic too. MW and ME sort them starting with (ا) marked with the vowel /i/, i.e., *Kasrah*, followed by (أ) marked with the vowel /a/, i.e., *Fathah*. This is the reverse of the order of short vowel diacritics with other letters. Python systematically follows the same order of vowel diacritics as all other letters. This stand is justified linguistically since in the corresponding long vowels *Alef* and *Ya* are in the same order.
- Sorting words with and without the definite morpheme {-ال}, which is a prefix attached to definite words, was variant across the three programs. ME and Python consider the first letter of the definite article {-ال} as an independent *Alef* letter and a basic constituent of a word. Accordingly, they arrange all the definite words starting with - {ال} before other words that do not contain it. Conversely, MW disregards {-ال} and sorts words based on the first letter following the prefix. This is a better stand than listing all definite words before others, especially in long lists.
- Since vowel diacritics are optional in Arabic (as well as nunation and germination symbols), many words are written without diacritics. The software programs were tested in sorting words with and without diacritics. MW and Python present the one without diacritics first, while in ME, the opposite happens. Either way is possible, but unification is essential. Furthermore, words without diacritics need not be considered independent entities since any word can be written in a diacritized or non-diacritized way.
- Regarding varying lengths of different words, the MW sorting system totally depends on the alphabetical order of the letters by comparing the first, second, and third letters of the word and so on. Varying vowel diacritics are disregarded in the sorting process. Conversely, ME and Python order words basically based on the vowel diacritics of the word regardless of its length. Python and ME sorting systems give more weight to vowel diacritics, while the MW sorting system gives more weight to letters. Clearly, considering the alphabetical order of letters with words of varying lengths is more reasonable, as letters are more phonologically salient than short vowel diacritics and including short vowel diacritics is optional in Arabic writing.
- The comparison list contains words with geminated letter. A geminated letter means a double-consonant or Prolonged consonant letter. In MW and Python, words with geminated letters precede words without geminated letters. However, the opposite order was found in ME’s generated list. Geminated letters are more phonologically salient, and accordingly, it is more coherent with sorting other characters to order it first.
- Comparing words that start with a lengthened (Prolonged) vowel /a:./ mark on the *Alef* letter, and a regular *Alef* as in أملا (I hope) and أملا (in hope of), MW and ME place regular *Alef* before Prolonged *Alef*. Python orders them in the opposite way.

Python's order is more linguistically justified since the Prolonged *Alef* actually comprises a double vowel of the same kind. A Prolonged vowel is longer and more salient phonologically, and it is more convenient to place it before the normal *Alef*. In this way, both doubled vowels and doubled consonants follow the same rule and are ordered before a single vowel and a single consonant.

Answers to Research Questions

Based on the above discussion, the answer to the first and second research questions is clearly “no.” The generated lists of the three programs differ in many aspects. The software programs lack a standardized method of alphabetical sorting of Arabic words accompanied by special characters. Neither MS- nor ME-sorted lists are identical to Python's list.

The answer to the third question, which requires a recommended *secondary* system for unification of alphabetical sorting of Arabic words with special characters across different software programs, can be achieved by referencing the linguistic features of Arabic letters and sounds. These include reference to the trilateral underlying roots of the words, phonological strength and salience, as well as the original alphabetical order of Arabic letters. Based on the examined special characters in the present study and the linguistic justifications mentioned above, the following *secondary* alphabetical order recommendations are proposed:

- The programs are recommended to sort *Alef* with glottal stop (ʾ) before *Alef* with the one without (ʰ) (i.e., *Hamzat Alqate'* before *Hamzat Alwasl*).
- Sorting of short vowel diacritics in Arabic (َ ُ ِ) is recommended as follows: *Fatha* /a/, *Dhamah* /u/ then *Kasrah* /i/ (including *Alef*).
- It is more logical to disregard the definite prefix {-ال} and sort words based on the first letter following the prefix. This order is better than listing all definite words before others, especially in long lists.
- When sorting words with and without diacritics, either order is possible. However, words without diacritics are the unmarked choice, and this can be a reason to begin with them first.
- Considering the alphabetical order of letters with words of varying lengths is more reasonable than sorting words based on their short vowel diacritics.
- It is more coherent with other suggested rules of ordering to sort geminated letters before similar non-geminated letters.
- A prolonged vowel is better sorted before a similar non-prolonged vowel.

Conclusion

The study revealed a lack of consistency in the alphabetically ordered lists generated for the same group of Arabic words containing special characters when utilizing two Microsoft Office 2021 programs, namely MW and ME. The list was ordered differently in Python. This is not just a simple technical error—it's a linguistic and cultural oversight with consequences for data integrity and accessibility. This result highlights a critical issue requiring solutions from software developers. Efforts should be made to harmonize the sorting criteria across programs regarding short vowel diacritics, definite and indefinite words, words with and without diacritics, and words with gemination and prolonged vowels. Program users must be cautious not to consider alphabetically ordered generated outputs of the programs as identical, especially when transferring lists between programs.

Results revealed that, except for nunnation and final vowel diacritics, all the investigated special characters are morphemic and have a semantic effect just like a letter. Accordingly, they must be given the status of a letter in computerized programs. This fact is important when electronic normalization of texts is required. Linguistic features of letters and special characters can also be used to establish a standardized system for alphabetically collating Arabic special characters. Some of these linguistic features include the underlying trilateral root form of the word and phonological salience of different characters. Finally, the study proposed a secondary alphabetical system for collating special characters in Arabic that aims to unify collation rules and is justified through linguistic features as a step toward a standardized, intelligent sorting algorithm that respects the structure of Arabic. Software developers are encouraged to consider the system when developing an updated version of their programs.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

No AI technology tool was used in the writing process.

Disclosure Statement

The author reports there are no competing interests to declare.

Availability of Data and Materials

The data used for analysis in this study can be accessed upon request through email communication with the corresponding author.

References

- Al-Onazi, B. B., Alotaib, S. S., Alshahrani, S. M., Alotaibi, N., Alnfai, M. M., Salama, A. S., & Hamza, M. A. (2023). Automated Arabic text classification using hyperparameter tuned hybrid deep learning model. *Computers, Materials & Continua*, 74, 5447-5465.
- Baroni, A. (2014). The invariant in phonology. The role of salience and predictability [PhD dissertation]. Universita Degli Studi Di Padova.
- Collation-Charts.ORG. (2024). <https://collation-charts.org/>
- Elmaghraby, A. S., El-Shihaby, S., & El-Kassas, S. (1989). Problems and peculiarities of Arabic databases, *Data Engineering*. 12(4), 2-11.
- Gussman, E. (2002). *Phonology: Analysis and theory*. Cambridge University Press.
- Hussain, D. S., & Karamat, N. (2003). Urdu collation sequence. *7th International Multi Topic Conference, 2003. INMIC 2003.*, 382-386.
- Motwakel, A., Al-onazi, B. B., Alzahrani, J. S., Marzouk, R., Aziz, A. S. A., Zamani, A. S., Yaseen, I. & Abdelmageed, A. A. (2023). Convolutional deep belief network based short text classification on Arabic corpus. *Computer Systems Science and Engineering*, 45, 3097-3113.
- Mustafa, S. (1996). Sorting problems of the standard Arabic character set. *Computer Standards and Interfaces*, 18(2), 159-173.
- Ryding, K. (2005). *A reference grammar of modern standard Arabic*. Cambridge University Press.
- Ryding, K. (2014). *Arabic: A linguistic introduction*. Cambridge University Press.
- Shihab, K. (2006). Arabic and multilingual scripts sorting and analysis. *Proceeding of 6th WSEAS International Conference on Applied Informatics and Communications.*, Elounda, Greece, 157–162.
- Street, J. (2020, June 11). *From A to Z - the surprising history of alphabetical order (text and audio)*. ABC News (ABC Radio National), Australian Broadcasting Corporation [Online]. Archived from the original on 2 July 2020. Street, Julie (June XX, 2020). <https://www.abc.net.au/news/2020-06-11/history-of-alphabetical-order-a-to-z/12320808>
- Wehr, H., & Cowan, J.M. (1996). A dictionary of modern written Arabic. Ithaca, NY: Cornell University Press.
- حلواني، محمد. (2017). *المغني الجديد في علم الصرف*. بيروت: دار الشرق العربي
- [Halawani, M. (2017). *A new comprehensive guide to Arabic Morphology*. Beirut: Arabic Easter Publishing House].

المخزومي، مهدي. و السامرائي، إبراهيم. (محقق) (1988) كتاب العين. (المؤلف: أبو عبد الرحمن الخليل بن أحمد بن عمرو بن تميم الفراهيدي البصري عام 170 هـ). بيروت: دار ومكتبة الهلال.
[Al-Makhzūmi, M. & Al-Samirā'ī, I. Al (Ed.), (1988). *Kitab al-'Ayn*. Beirut: Al Hilal Library].

Contact email: malamnea@imamu.edu.sa