Kurdish- Arabic Text in Computer with Non-Standards, Challenges and Recommendations

[1]Sardar O. Salih

[1] Department Of Computer Science, College of Science, University of Duhok, Kurdistan Region- Iraq

ABSTRACT

This paper determines challenges when using Kurdish-Arabic script (Kurdish and Arabic characters) with computer. Any written script in computer passes in three steps, started from input characters (creating characters), store created characters (encoding) and displaying characters on screen or printing on printer. In Kurdish-Arabic written system (Central Kurdistan) challenges occur on these three steps, as a result, there is not standard (compatible issue) for Kurdish-Arabic script, such as keyboard layout (position of characters and ordered on board) (input issue), encoding for storing script and fonts to displaying Kurdish-Arabic script when using Arabic based keyboard. This research tries to find issues with cases while using this script and proposed recommendations which help to reduce these challenges.

Keywords*:* Script, character, Unicode, Kurdish-Arabic characters, Encoding.

1. Introduction

To search specific keyword in Kurdish-Arabic characters .See Table 1 (Kurdish-Arabic characters) , it is difficult to predicate which written form for specific keyword to search, for instance if keyword (**weather**) is used for searching in English , there is not such other forms of characters to use despite it has two cases (small / capital letters), but in Kurdish such as (سەقا) means (weather) could be problem for text processing ,for instance (**سەقا / سەقا**), these two words are looks equal but not, due to, character ARABIC LETTER AE (ە) is not same encoding in two words (encoding issue) (G. H. Gautier, 1996) (Esmaili, 2012), will be explain later. The keyword could be written in Arabic pure characters such as (**سةقا**) this occur, as a result Arabic keyboard is used to generate data (input issue). Issues could be occurred when displaying script on screen , for instance, the keyword (**چیا**) is mean (mountain) in Kurdish is displayed (**چیا**) if (Zanst) Kurdish font is used (common user created) non-standard font but if standard font (Arial) is used to render, it looks like (**ضیا**) (misspelling/ displaying issue) (Shamsfard, 2011) (W3Team, 2017). These issues (challenges) with others will be mention in this research with some recommendations to reduce them.

Table 1 Kurdish-Arabic characters

Character	#	Character	#
ش	15	ئ	1
ع	16	ا	2
غ	17	ب	3
ف	18	پ	4
ڤ	19	ت	5
ق	20	ج	6
ک	21	چ	7
گ	22	ح	8
ل	23	خ	9
ڵ	24	د	10
ن	25	ر	11
و	26	ڕ	12
ۆ	27	ژ	13
		س	14

Academic Journal of Nawroz University
(AJNU) Volume 9, No 1 (2020).
Regular research paper : Published 4 March 2020
Corresponding author's e-mail : sardar.o@gmail.com

doi : 10.25007/ajnu.v9n1a584

2. Literature Review

In computer each character in the text is stored as unique number for example a, b, c, …etc. is stored as 97, 98, 99, …etc. respectively. For example, ASCII (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange) encoding system uses one byte of computer memory to save 256 characters, which can store English letters uppercase, lowercase, control character and some punctuations. The weakness of ASCII code cannot support non-English language such as Kurdish- Arabic characters (w3school, 2016) . Since most countries use characters exclude ASCII system .ISO International Standards Organization publishes (ISO 8859-1-1) to ISO (8859-1-15), ISO 8859-Jp and ISO 8859-KR to support most characters in the world including Arabic characters mentioned in (ISO 8859-1-6), but unfortunately, it supports only Arabic character and exclude some specific characters for some languages such as Kurdish , Urdu , Persian,…etc. ('ISO-8859-1 Reference', 2015). Thus, non-standard Kurdish fonts such as (Zanes and Ali) are used to display text with Kurdish specific characters which are not in Arabic ISO, such as (پ, گ, چ, ژ, …etc.) instead of some characters which are in Arabic ISO (not in Kurdish Language) like (ذ,ض, ث.. etc.). The following table shows Arabic ISO characters are not in Kurdish Language with corresponded Kurdish-Arabic character rendered when applying non-standard Kurdish fonts (G. Gautier, 1998).

Table 2 Kurdish Characters are not in Arabic ISO

Arabic characters rendered to corresponded Kurdish character when using non-standard Kurdish fonts	**Kurdish Characters are not in Arabic ISO 8859-1-6 and rendered to its corresponded Arabic characters**
ذ	ژ□
ض	چ□
ث	پ□
ط	گ□
ئ	ێ□
ؤ	ۆ□
ة	ه□
ظ	ڤ□

Example, when user type (ذ), it is rendered as (ژ) as show in Table 2 with Kurdish non-Standard font, this technic in some away help computer users to write document in Kurdish-Arabic characters , but this lead to challenge when document is moved to another computer with no installed Kurdish font , and when document is sent by email.

Recently , Unicode consortium comes with a system called Unicode, it provides unique number for almost all characters, punctuation and symbols in the word (Unicode, 2017). Fortunately, Kurdish- Arabic characters are supported on it. See table below Kurdish-Arabic characters with its Unicode. Despite Kurdish-Arabic characters are supported in Unicode standard and due to, there are not compatible systems, issues (problem) for text processing could raise.

Table 3 Kurdish-Arabic characters with its Unicode code by Unicode team KRG department of IT.

Unicode	Char	#	Unicode	Char	#
U+063A	غ	٢١	U+0626	ئ	١
U+0641	ف	٢١	U+0627	ا	٢
U+06A4	ڤ	٣٢	U+0628	ب	٣
U+0642	ق	١٢	U+067E	پ	٣
U+06A9	ک	٢٢	U+062A	ت	١
U+06AF	گ	٣٢	U+062C	ج	١
U+0644	ل	٣٢	U+0686	چ	٢
U+06B5	ڵ	١٢	U+062D	ح	٢
U+0645	م	١٢	U+062E	خ	٢
U+0646	ن	٢٢	U+062F	د	٣١
U+0648	و	٢٢	U+0631	ر	١١
U+06C6	ۆ	٢٢	U+0695	ڕ	٢١
U+0647	ه	٣٣	U+0632	ز	٣١
U+06D5	ە	١٣	U+0698	ژ	٣١
U+06CC	ی	٢٣	U+0633	س	١١
U+06CE	ێ	٣٣	U+0634	ش	١١
			U+0639	ع	٢١

4. Chalenges

doi : 10.25007/ajnu.v9n1a584

This paper state challenges occur when using Kurdish-Arabic characters in computer with cases and look up how these challenges can be proposed to solve.

4.1 KEYBORD Layout

There is not standard (unified) keyboard layout (key position and order on board) for Kurdish-Arabic characters as it available in most countries. As a result, raise to problems such as less experience, usability, compatibility. Since windows operating system is released, Kurdish users use Arabic based keyboard with non-standard Kurdish font to write their script on Kurdish-Arabic characters, as it mentioned, this cause to misspelling when document transfer from one computer to another (encoding and displaying issue). Most operating system builds keyboards for most counties in the world in their Operating system, unfortunately, Kurdish-Arabic keyboard are not included, Microsoft publishes tool call (Microsoft Keyboard Layout Creator) which helps to create keyboard is not as default from Microsoft Windows. thus, Kurdish keyboard is created using this tool, for instance (Talarsaz Keyboard, Kurdish Sorani, Badiny Keyboard, etc.). Recently, Microsoft Windows 10 adds default keyboard for Kurdish- Arabic characters named (Central Kurdish keyboard) see figure1 and figur2 Microsoft and user created keyboard are different in case position of characters and encoding. variety keyboard layout created for Kurdish-Arabic without unified keyboard which causes issues such as:

- Different keyboard layout with different character position and order causes user less experience to find position of character on it. For instance, Windows 10 Central Kurdish keyboard and User created Kurdish keyboard (Talarsaz) is totally different in case of character position and order see Figure 1 and Figure 2.
- Assigning Unicode encoding for each characters on board should be considered because Unicode encoding could have more than one encoding for one character have same shape, for instance Unicode ARABIC LETTER AE (ه) in Kurdish has two encoding (U+200C) or (U+06D5) and same shape in final position with space which one could be use? compatibility issues (Esmaili, 2012).
- Two characters (ك /ک) ARABIC LETTER KAF (ك) and ARABIC LETTER OPEN CAF (ک) have the same (shape) which of them to use on keyboard.

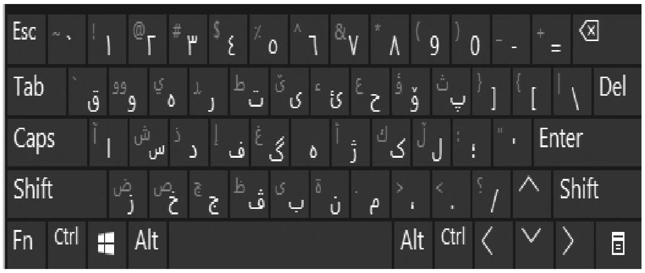

Figure 1 Central Kurdish Keyboard Windows 10 Default

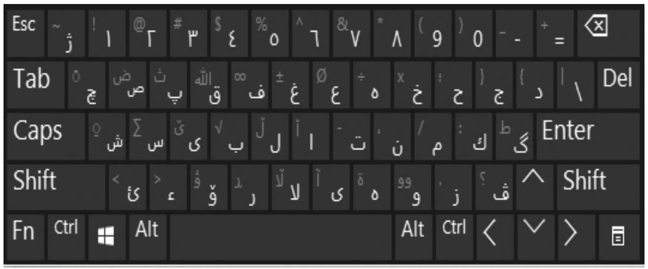

Figure 2 User Created Layout Talarsaz

4.2 Unicode and Kurdish-Arabic Characters

Assume unified keyboard is created for Kurdish-Arabic characters, the position of keys in keyboard are arranged as agreed, in this case Unicode encoding should be consider, it means which character is used with which Unicode Standard encoding. Difference encoding to same glyph (character) lead to processing text issue.

4.2.1 Case Study

Look at Microsoft Excel sheet, in Table 4. Word (سەقا) is created with two keyboards Microsoft and user common created without considering Unicode encoding, they look equal, but in Microsoft Excel sheet when they compared the result is not equal (false) (encoding issue). This occur due to UNICODE ARABIC LETTER AE, U+06D5(ە) and UNICODE ARABIC LETTER HA, U+0647 (ھ) can have same glyph (shape ە) when they are in final position with space. This because UNICODE ARABIC LETTER HA, U+0647 (ھ) has two pronunciations in the Urdu language , when its isolated and in final position of word its (pronounced e) and when it is in initial and middle of word its considered another word and (pronounced h), therefore, two Unicode encoding with same character are created (Esmaili, 2012) (G. Gautier, 1998).

Table 4 comparison of words contains (ە) in final position created with difference Unicode encoding

Using unicode ARABIC LETTER HA, U+0647 (ھ)	Using unicode ARABIC LETTER AE, U+06D5(ە)	IsEqual
سەقا	سەقا	FALSE
سەما	سەما	FALSE

In addition, Same issue with Unicode encoding for ARABIC LETTER KAF 0643 (ك) and ARABIC LETTER KEHEH 06A9 (ک) (QasemiZadeh, Rahimi, & Ghalati, 2005), look at table below, how the result of comparison is false and they are look equal. different Unicode encoding lead to Kurdish-Arabic text processing issue.

Table 5 comparison of word contains (ک) created with difference Unicode encoding

Using unicode ARABIC LETTER KEHEH 06A9 (ک)	Using unicode ARABIC LETTER KAF 0643 (ك)	IsEqual
کوردستان	كوردستان	FALSE
کەرکوک	كەركوك	FALSE

4.3 Transferring Data

When using Arabic based keyboard to write Kurdish-Arabic script with Kurdish font and transferring to another computer cause misspelling and unclear text. In this case all characters are in Kurdish language will be replaced with Arabic characters corresponded in Table 2. For instance, the character (ژ) is replaced with character (ذ) in Arabic.

4.3.1 Case Stydy

The following table shows Kurdish paragraph in left side which is created using Kurdish non-standard font, it is clear and readable but in the right side it rendered in computer with not installed Kurdish font, it is not cleared (misspelling).

Table 6 paragraph with Kurdish non-standard font and without Kurdish non-standard font

Using Kurdish font	Without Kurdish font
طشت ثيَنوسى ئةم دنياية دةست لةسةر سنط ذن ل ذياندا هاتنة بةردةم روخسارى تو ئةو ثيَنووسةى طول لة سةر	طشت ثيَنوسى ئةم دنياية دةست لةسةر سنط ذن ل ذياندا هاتنة بةردةم روخسارى تو ئةو ثيَنووسةى طول لة سةر

4.4 Using Arbic Keybord to Write Kurish Text

Due to Kurdish-Arabic keyboard is ignored in most Operation System (OS) such as Window, android and iOS. Thus, users use Arabic keyboard (characters) (see Figure 3) to write Kurdish text, as a result, the text is misspelling with excluded Kurdish characters in table 2.

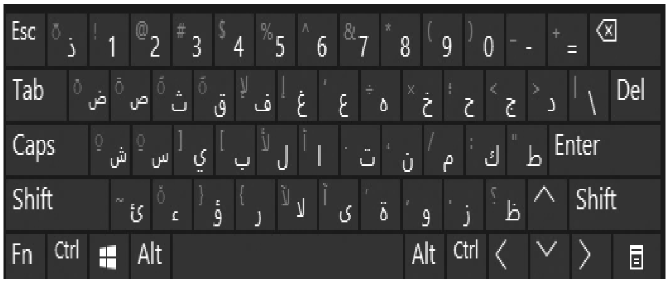

Figure 3 Arabic Keyboard Layout

4.4.1 CASE STUDY

The following text is written used Arabic keyboard. It can be read in some away but it is misspelling (grammar errors). In this case, it is difficult to use in computer processing such as finding, sorting, etc. For instance, if to try search (دلوڤان) means (mercy) in Figure 4, which keyword to use (دلوفان) or (دلوڤان) to find it. See Figure 4 the gray highlighted characters are Arabic characters not in Kurdish language should be replaced with Kurdish characters to be clear and correct in spelling. See Figure 5.

بنافئ خودايئ مةزن ودلوفان
ئةف تئكسته هاتية نفئسين ب كيبوردئ عة ره بى

Figure 4 Kurdish text uses Arabic Keyboard

بناڤێ خودایێ مەزن ودلوڤان
ئەڤ تێکسته هاتیه نڤێسین ب کیبوردێ عه رەبی

Figure 5 Kurdish Text after corrected

5. Discussion and Recommendation

If look at both Keyboard layout figure1 and figure2 and look at position of characters and compared with the Arabic keyboard Figure 3, the user created keyboard is near to Arabic keyboard in case character position on board but Microsoft windows build-in Keyboard is totally different in case character position on board. These two layouts with variety users created layout cause confusing (headache) to user which one to use. The keyboard is near to Arabic keyboard layout is best choose in case of usability (easy to use) for most users, due to, Arabic keyboard is used frequently and available on most operating system such as windows, Android, etc.

Assume keyboard layout is chosen, the Unicode encoding should be considered such as the characters UNICODE ARABIC LETTER AE, U+06D5(ە) and UNICODE ARABIC LETTER HA, U+0647 (ه) cause issue when using in final position followed space, both looks same. See Table 4, this cause text processing such as comparison, sorting and searching (different Unicode encoding). this case UNICODE ARABIC LETTER HA, U+0647 (ه) *should not be* used in final form and it is used as its in form of (ه) in Kurdish text (G. Gautier, 1998). In addition, ARABIC LETTER KAF 0643 (ك) and ARABIC LETTER KEHEH 06A9 (ک) both looks same in text see Table 5 (different Unicode encoding), one of them should be used. In this case ARABIC LETTER KEHEH 06A9 (ک) is Kurdish and Persian character and not Arabic, *this should be* used and avoid using ARABIC LETTER KAF 0643 (ك).

If keyboard layout and prepare Unicode encoding are chosen, *this could be* useful to avoid Kurdish users to write text using Arabic keyboard see Figure 3 (common written especially with Mobile users) which excludes Kurdish characters, also this avoid use Arabic keyboard with Kurdish font which causes misspelling text when transferred from one computer to another.

6. Conclusion

As it has been mentioned, to avoid challenge when processing, transferring and storing Kurdish-Arabic characters in computer, unified keyboard layout requires with characters are located on based in (Arabic keyboard layout) with simply changing characters are not in Kurdish language with Kurdish characters not in Arabic as mentioned. Then, Unicode encoding should be assigned with concerning characters such as (ە) and (

ـه) and avoid using (ـه) in final position followed space and instead of it using (ه). As well as, character (ک) should be used on keyboard and avoid using character (ك) which lead to ambiguity. After assigned keyboard layout with its Unicode encoding as it supposes, then, this could be help user to avoid using Arabic keyboard with Kurdish font and Arabic keyboard without Kurdish font (common on mobile devices) which raise misspelling and unclear text.

7. References

1. Esmaili, K. S. (2012). Challenges in Kurdish text processing. *ArXiv Preprint ArXiv:1212.0074*.
2. Gautier, G. (1998). Building a kurdish language corpus: An overview of the technical problems. *Proceedings of ICEMCO*. Retrieved from http://ggautierk.free.fr/e/icem_98.htm
3. Gautier, G. H. (1996). Dirêjî Kurdî : a lexicographic environment for Kurdish language using 4th Dimension®. In C. University (Ed.), *5th International Conference and Exhibition on Multilingual Computing (ICEMCO)* (Vol. 5, p. Session of the 12th April). Londres, United Kingdom. Retrieved from https://hal.archives-ouvertes.fr/hal-00676294
4. ISO-8859-1 Reference. (2015). Retrieved 7 October 2017, from https://www.w3schools.com/charsets/ref_html_8859.asp
5. QasemiZadeh, B., Rahimi, S., & Ghalati, M. S. (2005). Challenges in Persian Electronic Text Analysis, 5.
6. Shamsfard, M. (2011). Challenges and open problems in Persian text processing. Proceedings of LTC, 11.
7. Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing (TALIP), 8(4), 14.
8. Unicode. (2017). Unicode. Retrieved 6 October 2017, from http://www.unicode.org/standard/WhatIsUnicode.html
9. w3school. (2016). HTML Character Sets. Retrieved 7 October 2017, from https://www.w3schools.com/charsets/default.asp
10. W3Team. (2017). CSS Web Safe Fonts. Retrieved 30 June 2018, from https://www.w3schools.com/cssref/css_websafe_fonts.asp

doi : 10.25007/ajnu.v9n1a584