Journal of Chinese Language and Computing 15 (4): (211-218)

Multilingual Hybrid Text Processing in Ancient Uighur (Chaghatai) Digitalized System

Dilmurat Tursun
College of Information Science and Engineering,
Xinjiang University, Urumqi, Xinjang 830046, P.R. China
E-mail: dlmlttu@xju.edu.cn

Abstract

This paper analyzes and presents the Unicode standard codepage, keyboard layout, implementation of Uighur and Chaghatai (ancient Uighur script) IME for Chaghatai Digitalized Processing System and Multilingual Processing[1].

Keywords

Chaghatai Language, Digitalized Processing, Standard, Unicode, Keyboard Layout

1. Introduction

This research mainly considers and discusses system codepage in special techniques to multilingual processing of ancient Uighur literatures (Chagatai for abbreviation in the following text). Based on detailed analysis to Arabic code page, Farsi codepage and Uighur codepage in Unicode standard, we presented a codepage and keyboard layout, which is compatible with Chaghatai, Arabic, Farsi, Uighur and Latin characters, is proposed. It is a key technique for achieving specialized Chaghatai word processing systems.

2. Unicode codepage for Chaghatai script

In the year 2002, the Information Science and Engineering college of Xinjiang University successfully developed a Uighur language processing system based on Unicode 4.0 standard, which works on Microsoft Windows 2000/XP platform, and gathered valuable experiences about processing Arabic Unicode charsets. At the same time application for Uighur Unicode code page was sent to ISO.

[1] Sponsored by Chinese National Natural Science Foundation (NO.60363003)

With the achievements and experiences in developing Uighur platform, considering compatibility with Uighur, we think Chaghatai language codepage should satisfy the following requirements:

1. Must follow Unicode standard.
2. Formats and writing rules of selected characters must comply with those of the Chaghatai script.
3. Must be compatible with Arabic, Farsi, Urdu and Uighur.
4. Changing old characters or adding new characters are not allowed. They must be located in Unicode basic area when stored or processed.

Chaghatai alphabet consists of 33 characters. Based on the shape of Chaghatai characters appeared in ancient literatures and the technical requirements proposed above, a draft for Chaghatai codepage table is as in Table 1.

7	6	5	4	3	2	1
چ 0686	ج 062C	ث 062B	ت 062A	پ 067E	ب 0628	ا 0627
14	13	12	11	10	9	8
ژ 0698	ز 0632	ر 0631	ذ 0630	د 062F	خ 062E	ح 062D
21	20	19	18	17	16	15
ع 0639	ظ 0638	ط 0637	ض 0636	ص 0635	ش 0634	س 0633
28	27	26	25	24	23	22
م 0645	ل 0644	گ 06AF	ک 06A9	ق 0642	ف 0641	غ 063A
35	34	33	32	31	30	29
ء 0621	ئ 0626	نگ 06AF+0646	ی 06CC	ھ 0647	و 0648	ن 0646
			39	38	37	36
			ة 0629	ؤ 0624	أ 0623	آ 0622

Table 1. Chaghatai code page for Unicode 4.0

Although there are 33 characters in Chaghatai alphabet, in many literatures there appeared many words and sentences cited from Quran. To correctly display these sentences we also added characters between No 30 and No 39.

To recognize and locate Chaghatai characters correctly, we collected and analyzed a great deal of Chaghatai literature and the different shapes and compound rules of Chaghatai characters in those literature, found out the corresponding relationship between Arabic Unicode basic area and 3 extended areas. For example:

1. The usage and pronunciation of Chaghatai character ى is the same as those of Uighur character 0649. But another two formats of character 0649 are different from formats of ى .So we used Farsi character 06CC ى for whose all formats are the same as Uighur character ى .
2. The shape of Chaghatai character ھ is very similar to Uighur character 06BE. But the character 06BE only has two different forms, while Arabic character 0647 has four different forms, which are the same as Chaghatai character ھ. So the character 0647 is used for Chaghatai character ھ.
3. Different forms of character 06A7 are the same as Chaghatai character ڧ . But automatic shape selection process of character 0641 is easier and its shape is similar to ڧ. Chaghatai character 0623 and character 0672 also has such relations.
4. In Unicode table we could not find any character similar to character No 33. Considering its shape and different forms we decided to solve it with composite format (0646+06AF).

Uighur Characters	Uighur Characters
ى 0649	ى 06CC
ھ 06BE	ھ 0647
ٵ 0672 Non-Uighur Character	أ 0623
ۋ 06A7 Non-Uighur Character	ف 0641

Table 2. Additional Characters

3. Keyboard Layout

At present there is no a reference material or data about Chaghatai wordlist and frequency of the words that appear in the language. Because of that, we referred to Arabic, Farsi and modern Uighur language keyboard layouts when arranging Chaghatai keyboard. Figure 1 and Figure 2 show the arranged Chaghatai Keyboard Layouts.

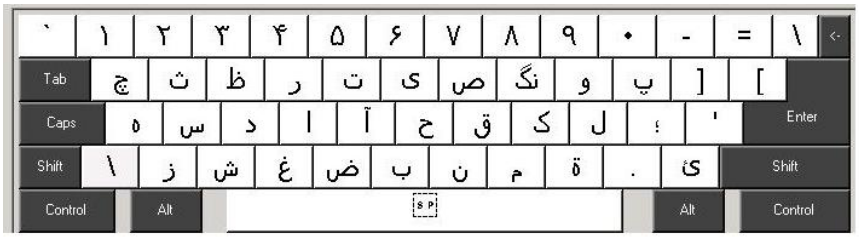

Figure 1. Chaghatai Keyboard Layout

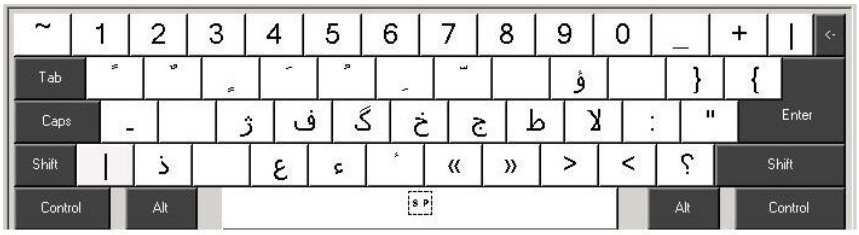

Figure 2. Chaghatai Keyboard Layout when Shift is pressed

The main considerations are:

1. As numbers in Chaghatai are Arabic numbers, keep their original locations.
2. Typists of Chaghatai scripts are usually accustomed to Uighur keyboard layout and there is a few in numbers. So keep the locations of the common characters the same as on Uighur keyboard.
3. As to some characters are different from Uighur characters, they are located at the areas easy to be remembered. For example: the location of character غ, which is common to Uighur and Chaghatai keeps unchanged. But character ع is located in Shift position of character غ. ۇ is located in Shift position of و. The different formats and pronunciations of Chaghatai character ك are the same as those of Uighur character ك. So they are located on the same key. Chaghatai character ذ is located in Shift position of ز. To make them easy to remember, غ and ع, ا and آ, arranged together.
4. Locations of characters that used to display sentences cited from Quran are kept the same as in
5. Arabic keyboard layout.
6. In Chaghatai script character نگ is composed of two codes. But in fact it is one character ڭ, so it is located as a composed character in the same key as Uighur character ڭ whose pronunciation is similar to نگ.

4 .Chagatai multi language processing Implementation Algorithm

Based on the Unicode Uighur (Chaghatai) codepage and keyboard layout, we developed a compatible processing system ChaghWord, which processes contemporary Uighur, Chaghatai, Latin and phonetic symbols. This system is able to input, display and print Arabic, Contemporary Uighur and Chaghatai and can corrects contemporary Uighur, Chaghatai with limited word list. It has the intelligent input and self-study functions and can generate two different phonetic symbols for Chaghatai. (For local users and Turkish phonetic symbols).

4.1 Chaghatai input method and algorithm

Through hook function, get the key value of the Arabic input, and change it according to the corresponding relationship and implement the Chaghatai input. The hook function of the chagh.dll is shown on figure 3.

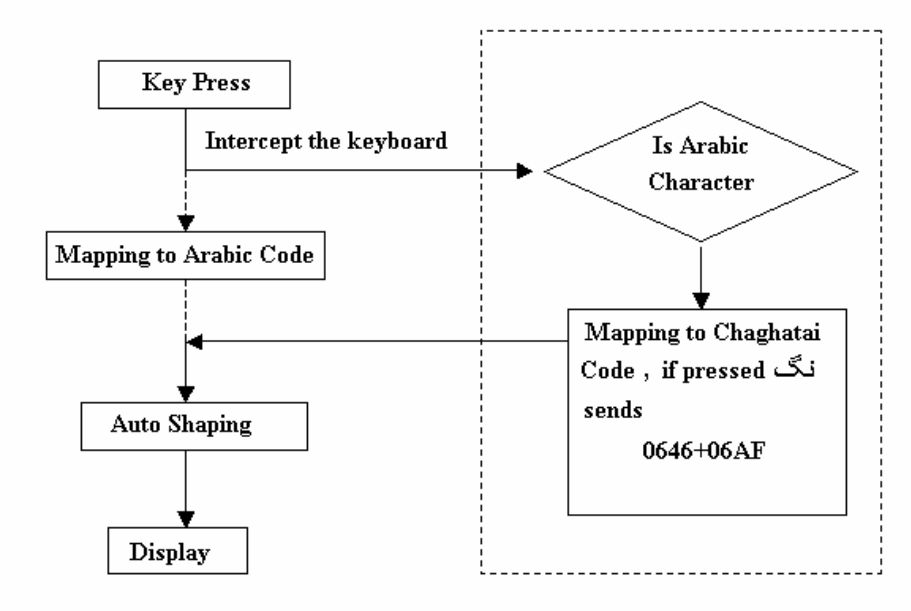

Figure 3. Ancient Uighur(Chaghatai) Input Implement Flow

4.2 Contemporary Uighur Input Implement Algorithm

When we used the Delphi 6.0/7.0 ,C++ Builder 6.0, Power Builder 8/9 to process contemporary Uighur, we found that Arabic codepage can't meet the process of the contemporary Uighur. Therefore, we changed the Arabic codepage file cp_1256.nls, without affecting the old system. Through patching, added the 7 contemporary Uighur characters which are not found in the Arabic codepage to the codepage. The Unicode versions before 4.0, define the Uighur Character 06D5 wrongly as 06D5; AE; U; <no shaping>.

But this character is the right linked character in contemporary Uighur script. In the Unicode 4.0 version character-shaping table ArabicShaping-4_0_0, this character is re-defined as 06D5; AE; Right Joining; TEH MARBUTA. Therefore, we imposed the forced shape selecting to the character 06D5. The flow chart of the functions of the Uqur.dll is shown on figure 4 (in the dashed lines)

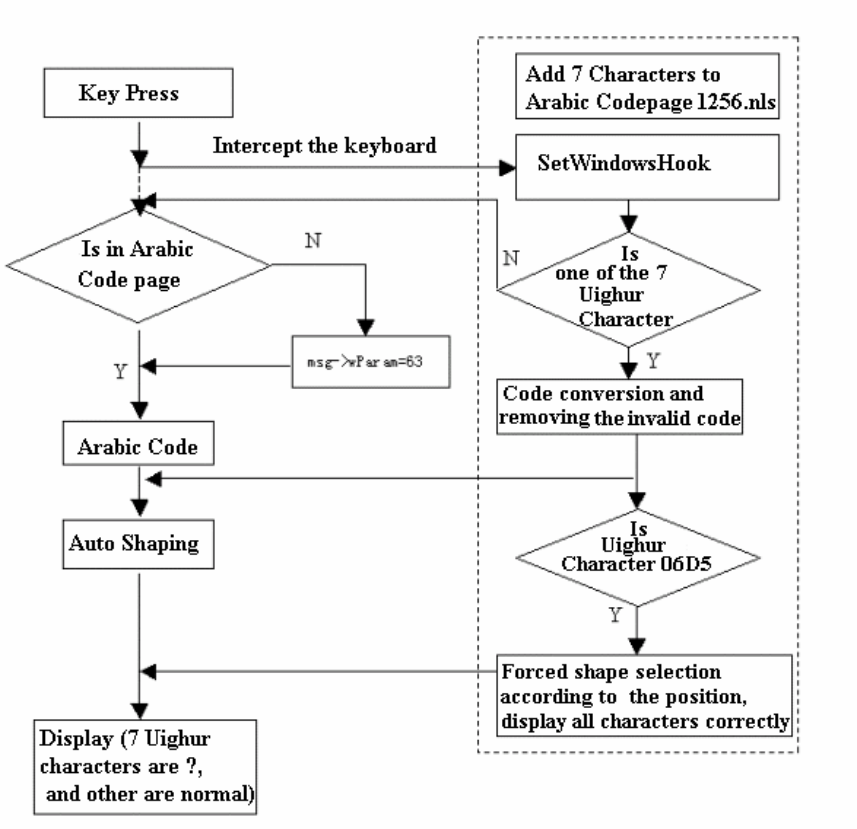

Figure 4. The flow chart of the hook function

5. Create Chaghatai Fonts

With the prerequisite of guaranteeing the normal function of the operating system, we made 4 dedicated Open type fonts, which are compatible with Arabic, Farsi, Urdu and contemporary Uighur. We used FontCreator, FontGrapher ,softy and other tools to create the fonts. We used Microsoft Volt for Font scripts. In the system fonts like Arial, Times New Roman, Tahoma, Microsoft Sans Serif some Uighur characters and their different shapes are not given. For some other letters, their shaping rules are different. Through patching Tahoma, Microsoft Sans Serif fonts and changing their script, we correctly displayed the Uighur and Chagatai characters on development tools and system interface.

6. Processing the existing Uighur and Chagatai materials

Formerly published Chagatay books and other materials are inputted and saved in two different formats. One is the Beida Fangzheng, Sanli publishing systems in DOS system. Another is a different non-standard input tool on Windows. Arabic Windows systems support automatic shaping. Therefore they use Unicode Basic area to save and use Unicode extended area to display. But The Beida Fangzheng publishing system uses different code for the different shapes of the same character when display and save. These are shown in Table 3.

Character Postion	Mid , Final, Initial , Isolate	Character Postion	Mid , Final, Initial , Isolate
Display 北大方正 QuWei Code	ن ـنـن 9048,9047,8947,8948	Display (Windows,Unicode)	ن ـنـن FEE6,FEE8,FEE7,FEE5
Mem 北大方正 QuWei Code	9048,9047,8947,8948	Mem (Windows,Unicode)	0646,0646,0646,0646
Storage 北大方正 QuWei Code	9048,9047,8947,8948	Storage (Windows,Unicode)	0646,0646,0646,0646

Table 3. Uighur and Chaghatai characters in DOS Beida Fangzheng and Windows

To use those Chagatay materials inputted in DOS system with Beida Fangzheng format, we developed bi-direction conversion program between DOS Beida Fangzheng format and Windows Unicode format. The working flow is shown in Figure 5.

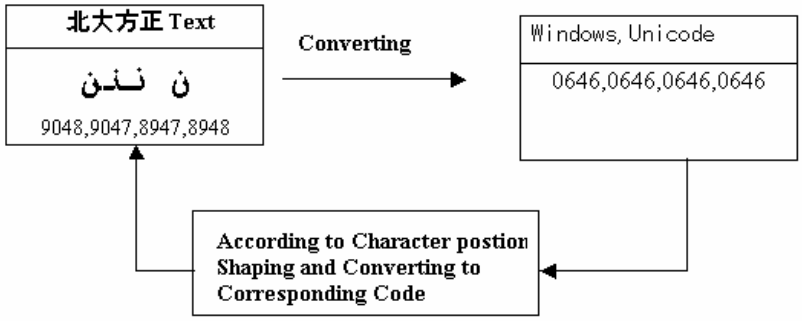

Figure 5. Data flow between Dos Beida Fangzheng format and Windows Unicode format

The other Chagatai documents inputted with not-standard input methods need to be converted after code analysis. We also developed the convert program between standard Unicode and Uighursoft Alkatip, Almas Office, Coltegen Silk Road 2000 and other input methods.

7. Summary

The first task in Chaghatai Digitalized Processing system is to finish standard Chaghatai codepage. The developed Input model for Chaghatai dictionary using research results from this paper is functioning stably, showing good compatibility with Arabic, Farsi and modern Uighur. It proved that the codepage and keyboard layout has practical values in setting standard Unicode code for Chaghatai script and it is possible to send it to ISO as a part of Uighur Unicode codepage.

Input method developed in accordance with the keyboard layout that proposed in this paper proved We developed the ChaghWord system based on the this input method and the keyboard layout that proposed in this paper proved to be practical, easy-to-use, and is welcomed by Chaghatai language experts. It can be released as the part of standard Uighur input method.

This paper is our recent approaches to codepage and keyboard layout for Chaghatai script which is compatible with Arabic, Farsi, and Contemporary Uighur, based on the detailed analysis on the Arabic, Farsi, and contemporary Uighur codepages on Unicode standard. This is our initial experiment and any ideas and questions are welcome.

References

[1] Nadine Kano, “Introduction to International software development on Windows 95 Windows NT”, “translated by Xia lili, Tsinghua University Press, 1998.3

[2] Mamattursun, Baodun, “Chaghatai Language Dictionary”, XinJiang People’s Press, 2002.3

[3] Ablimit Ahat, “Chaghatai Uighur language”, Xinjiang University Press, 2003.7

[4] “Uighur Star: Uighur-Chinese Windows operating system platform technical report”, “ project evaluation material”, Xinjiang University EE department, 1997

[5] Wulamu,”Variorum of Uighur Ancient Literature Words and Expressions”, National Press, 1995.6

[6] Niu ruji , “Introduction to Uighur ancient writing and literature”, Xinjiang People’s Press, 1997.4