Research on Unicode and Keyboard Layout for Chaghatai Digitalized Processing System*

Dilmurat.Tursun Wayit.Ablez Turgun.Ibrahim
Institute of Information Science and Engineering,
Xinjiang University, Urumqi, China, 830046
E-mail: dlmlttu@xju.edu.cn

Abstract*: This paper analyzes and presents the Unicode standard codepage, keyboard layout for Chaghatai(ancient Uighur script) Digitalized Processing System.*

KeyWords*: Chaghatai Language, Digitalized Processing, Unicode Standard, Keyboard Layout*

1. Introduction

This research mainly considers and discusses system codepage in special techniques to multilingual processing to ancient literatures. Based on detailed analysis to Arabic codepage, Farsi codepage and Uighur codepage in Unicode standard, a multi-purpose code page blueprint, which is compatible with Chaghatay , Arabic, Farsi ,Uighur and Latin characters, is proposed. It is a key technique for achieving specialized word processing systems.

2. Choosing Unicode codepage for Chaghatay script

In the year 2002, the Information Science Institute of XinJiang University successfully developed a Uighur language processing system for information exchanges 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 codepage was sent to ISO.

With the achievements and experiences in developing Uighur platform, considering compatibility with Uighur, we think Chaghatay language code page must satisfy the following requirements:

1. Must fit Unicode standard.
2. Formats and writing rules of selected characters must be the same with those of the Chaghatay 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.

Chaghatay language alphabet consists of 33 characters. Based on the shape of Chaghatay characters appeared in ancient literatures and the technical requirements proposed above, a draft for Chaghatay code page table is as follows:

7	6	5	4	3	2	1
چ 0686	ج 062C	ث 062B	ت 062A	پ 067E	ب 0628	ا 0627

*Sponsored by Chinese National Natural Science Foundation（NO.60363003）。

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	ۇ 0624	أ 0623	آ 0622

Table 1 . Chaghatay code page for Unicode 4.0

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

When making code page for Chaghatay language, we discovered that some characters are in similar shape. Examples are as shown in Table 2.

Uighur Characters	ی 0649	ھ 06BE	ۈ 0672 (not Uighur character)	ڧ 06A7 (not Uighur character)
Chaghatay Characters	ی 06CC	ھ 0647	أ 0623	ف 0641

Table 2. Codes with simlar shape

To recognize and locate Chaghatay characters correctly, we analyzed and collected their different formats and composite rules under various conditions and found out their corresponding relationship with Unicode Arabic characters in basic area and 3 extended areas.
For instance:

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

3. Keyboard Layout

At present there is no a reference material or data about Chaghatay wordlist and frequency of the words that appear in the language. Because of that, we refered to Arabic,Farsi and modern Uighur language keyboard layouts when arranging Chagatay keyboard. The following are keyboard layouts of those three languages:

Figure A. Arabic Keyboard Layout

Figure B. Arabic Keyboard Layout when Shift is pressed

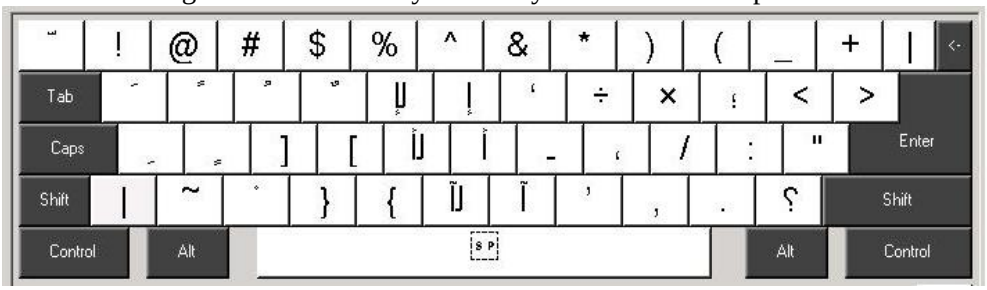


Figure C. Farsi Keyboard Layout

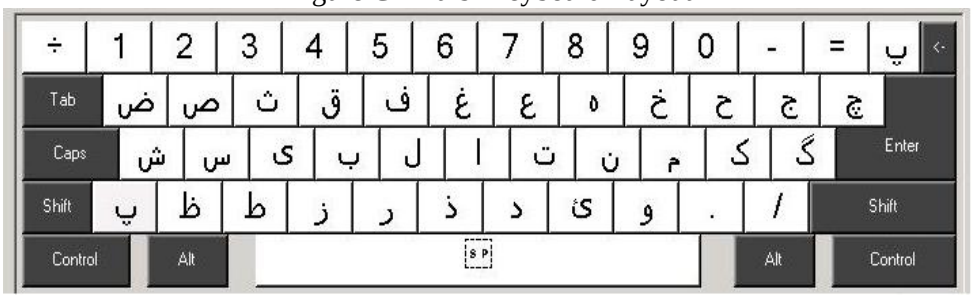


Figure D. Farsi Keyboard Layout when Shift is pressed

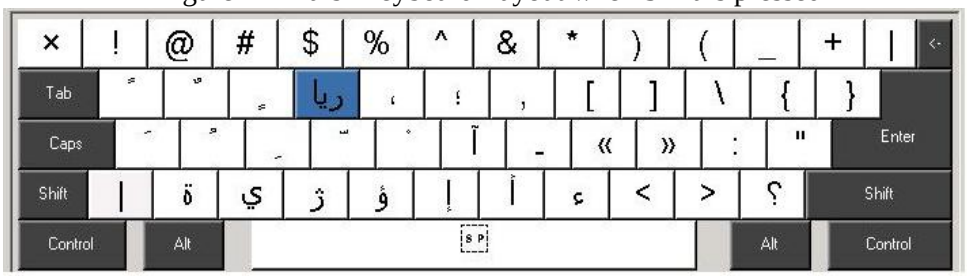
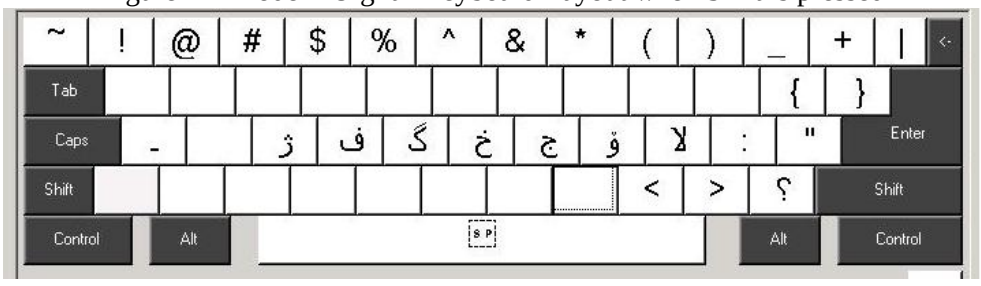


Figure E. Modern Uighur Keyboard Layout

Figure F. Modern Uighur Keyboard Layout when Shift is pressed



Arranged Chagatay Keyboard Layouts are as follows:

Figure G：Chagatay Keyboard Layout

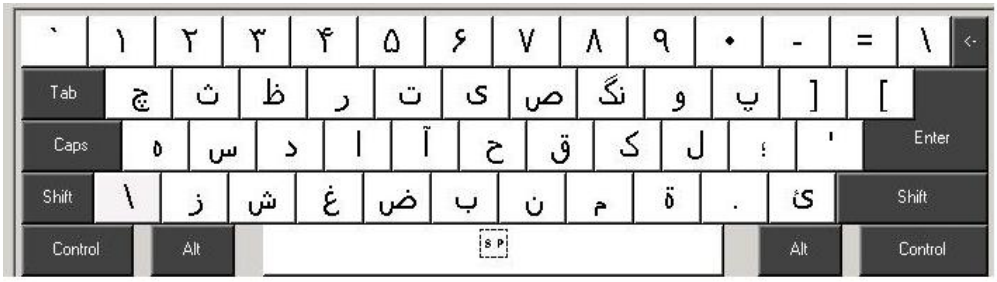
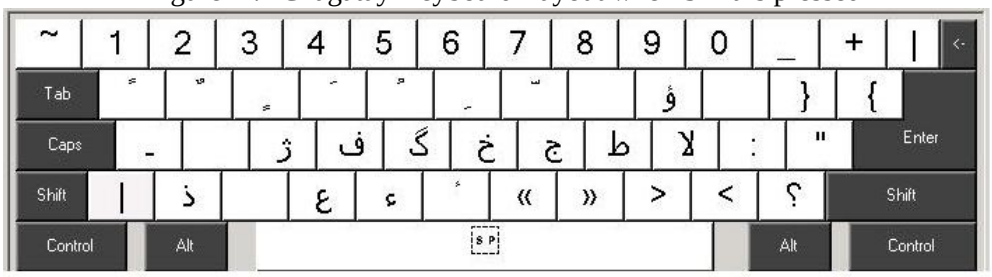


Figure H：Chagatay Keyboard Layout when Shift is pressed



The main considerations are:

1. As numbers in Chaghatay are Arabic numbers, keep their original locations.
2. Typists of Chaghatay scripts are usually accustomed to Uighur keyboard layout and there is a few in number. 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 Chaghatay keeps unchanged. But character ع is located in Shift position of character غ. ۇ is located in Shift position of ۋ. The different formats and pronunciations of Chaghatay character ك are the same as those of Uighur character ك. So they are located on the same key. Chaghatay character ذ is located in Shift position of ز.
4. Locations of characters that used to display sentences cited from Quran are kept the same as in Arabic keyboard layout.
5. In Chaghatay script character نگ is composed of two codes. But in fact it is one character, so it is located as a composed character in the same key as Uighur character ڭ which's pronunciation is similar to نگ .

4. Summary

The first task in Chaghatay Digitalized Processing system is to finish standard Chaghatay codepage. The developed Input model for Chaghatay 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 Chaghatay script and it is possible to send it to ISO as a part of Uighur Unicode codepage.

Input method developed in accordance with keyboard layout that proposed in this paper proved to be practical, easy-to-use, and is welcomed by Chaghatay language experts. This paper is our recent

approaches to codepage and keyboard layout for Chaghatay script, any ideas and questions are welcome.

Reference

Nadine Kano [US] ,translated by Xialili,Introduction to International software development on Windows 95 Windows NT ,Tsinghua Universtiy Press.

Mamattursun ,Baodun, Chaghatay Language Dictionary,XinJiang People's Press.

Ablimit Ahat, Chaghatay Uighur language, XinJiang Universtiy Press.

Introduction to authors:

1. Dilmurat Tursun (1956 --),graduated from physics science department of XinJiang University in 1983 ,Professor. Main research area: electronics, information management
2. Wayit.Ablez (1975 --),graduated from telecommunication department of LanZhou University, lector. Main research area: Multilingual information processing.