Urdu Component Development Project

Urdu Collation Application v1.0

October 26, 2007

CENTER FOR RESEARCH IN URDU LANGUAGE PROCESSING
NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES, LAHORE
PAKISTAN

Table of Contents

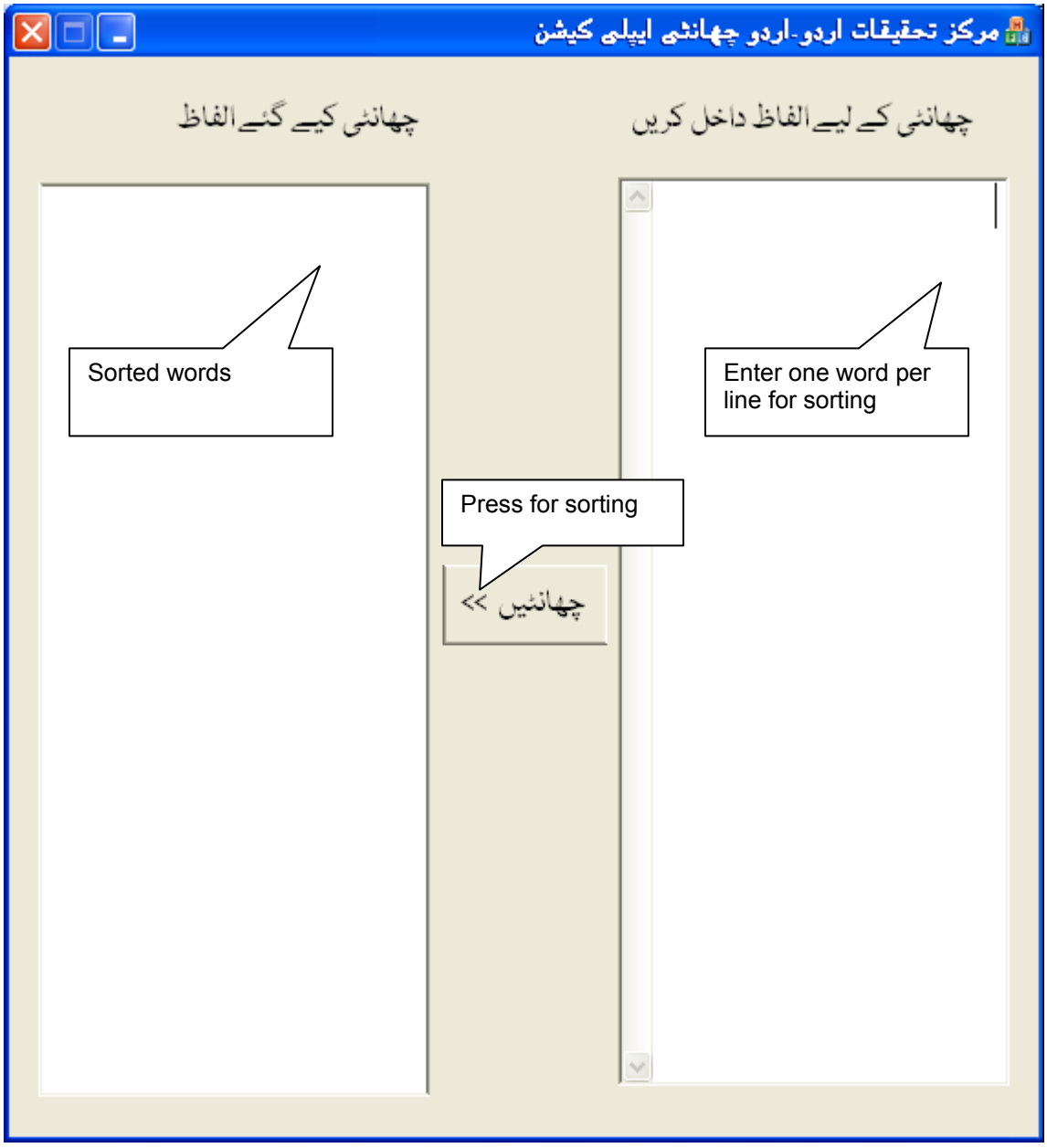
Revision History

Name	Change Date	Version	Description of Changes
Atif Gulzar	26-10-2007	1.0.0.0	Initial document

1 Introduction

Urdu collation application provides a language sensitive sorting for Urdu strings. This document enlists Urdu Collation application features and usage.

2 Application Usage



3 File Formats

3.1 Urdu Collation Table

This file contains the collation sequence of Urdu alphabet. Urdu sorting requires at least three levels. At the first level of sorting, only the basic Urdu characters will be sorted. Once the characters determine word sequence, aerab are used to determine the sequence of words having the same characters. Finally the third level is used to sort special symbols e.g. honorific mark.

The first line of this file contains the number (in hexadecimal) of collation elements (excluding composite collation elements). The remaining each line has format [Code][C_1,C_2,C_3] for singe character or [Code1,Code2] [C_1,C_2,C_3] for composite characters. Where *Code* is the Unicode value of the character and C_1, C_2 and C_3 define the level of collation sequence at level1, level2 and level3 respectively. The values of C_1, C_2 and C_3 must be less than 0x7F.