Urdu Collation Sequence

Dr. Sarmad Hussain, and Nayyara Karamat
Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences
Emails: sarmad.hussain@nu.edu.pk, mscs060@nu.edu.pk

Abstract: *This paper presents problems in Urdu collation sequence and proposed solutions by experts of Urdu. Key issues regarding Urdu collation sequence on which the major dictionaries differ are presented and a unanimous solution by the experts on these issues is also presented. This effort will make possible a standard Urdu collation sequence and make advance data processing possible for Urdu.*

Keywords: *Urdu standardization, Collation sequence*

1. INTRODUCTION

Achieving effective storage and retrieval of data in computers requires data to be stored using some standard encoding scheme and in some pre-determined order To achieve objective of efficient retrieval ascending or descending order is used for numeric data and lexicographic order is used for textual data. When storing, sorting and accessing textual data, the lexicographic order has to be well defined. This has been done for English and many other languages, where computational support has been available for a long time. With multi-lingual computing growing rapidly and more languages being incorporated into computing platforms, collation sequences of these languages also need to be determined. Urdu is one of these languages. Encoding scheme for Urdu characters has been addressed through the development of UZT (Urdu Zabta Takhti) [1] [2] and eventual upgradation of Unicode standard. In this paper problems in Urdu collation sequence are discussed and the respective recommended solutions are presented which were proposed by experts (see Appendix A) in a seminar conducted by URLSDF (Urdu and Regional Languages Software Development Forum) of Ministry of IT in collaboration with National Language Authority of Pakistan on 7th July, 2003.

2. URDU COLLATION SEQUENCE

Fundamental source of identifying the traditional collation sequence of any language can be its dictionaries. Numerous dictionaries are available for Urdu (e.g. [4, 5, 6, 7, 8, 9, 10 and 11]). This fact implies that to a large extent, lexicographic order has been defined for Urdu. However, closer analysis of Urdu dictionaries, most of which are hand-written or typed using type-writers, reveals that there is some variation between lexicographic orders of different dictionaries. Following is detail of problems in collation sequence of Urdu.

2.1 Points of Disagreement

Analysis of eight different dictionaries reveals that though most of Urdu characters follow an agreed sequence, the dictionaries do not agree on status and/or sequence of some characters. Details of issues related to these characters are given in the sub-sections below.

2.1.1 آ (Alif Mad)

Dictionaries do not agree on the status of آ as a character, separate from ا. Dictionaries which consider آ as a separate character, do not agree on whether it comes before or after ا. Thus, the following variation in word order exists in various Urdu dictionaries listed. See Reference section for the names of the dictionaries against the abbreviations listed.

Table 1. variation in sequence of ا and آ

آ not a separate character from ا	آ separate character, comes before ا	آ separate character, comes after ا
STCD, UHE, FA, NL	FLJ	FT, JUL, UL
ا	آ	ا
آ	آب	اب
آب	آپ	ایوان
آپ	ا	آ
اب	اب	آب
ایوان	ایوان	آپ

For the dictionaries that do not consider آ as a separate character, it is considered equivalent to اا (two Alifs) and thus words starting with آ come between ا and the word اب. For the dictionaries, which consider آ as a separate character, the word order changes completely, depending on whether آ comes before or after ا.

2.1.2 ں (Noon Ghunna)

It is not clear whether ں is an independent character or if it is a variation of ن. No word starts with ں and the letter only nasalizes the previous vowel, not adding any phonemic content by itself. However, dictionaries disagree whether it comes before or after ن. Thus, the following variation in word order exists in various Urdu dictionaries listed. See Reference section for the names of the dictionaries against the abbreviations listed.

Table 2. variation in sequence of ن and ں

ں before ن	ں after ن
FLJ, FT, STCD, NL, FA, UHE	JUL, UL
ماں	مان
مان	ماں

2.1.3 ھ (Do Chashmey Hay)

ھ is possibly not a character but a consonant modifier to indicate aspirated consonants. It combines with a variety of Urdu consonants to give new consonants, e.g. ب + ھ = بھ or پ + ھ = پھ. However, some dictionaries consider the combined character as a separate character, while others consider it as not a combined separate character but a sequence of ب ھ, resulting in the following word orders.

Table 3. variation in sequence of بھ and ب

بھ separate character	بھ sequence of (ھ ب ھ) before ہ)	بھ sequence of (ھ ب ھ) after ہ)
FT, JUL, UL	STCD	FLJ, UHE, FA, NL
باپ	باپ	باپ
بہن	بھابی	بھابی
بہنگی	بہن	بہن
بیٹا	بھنگی	بھنگی
بھابی	بہنگی	بھنگی
بھنگی	بیٹا	بیٹا

In the dictionaries, where بھ is considered as a combination of ب ھ the order of ھ vs. ہ is not clearly defined. In the first pass, the two are considered same; however, where exactly the same pattern exists, and the word order has to be decided only on the basis of

these two characters, the second pass either places ھ before or after ہ.

2.1.4 ۃ (Gol Tay)

It is not clearly defined in any dictionary whether ۃ is a variation of ہ or ت. Its position with respect to these characters is also not defined.

2.1.5 ے/ی (Chhoti and Bari Yay)

Though it is fairly conventional and agreed upon by all that ی comes before ے, various dictionaries do not agree whether the difference in collation is in the first or second pass of sorting, giving the following variations. In addition, it is not clear whether middle yay has to be distinguished as ی or ے, or the distinction collapses to a single middle-yay for sorting purposes.

Table 3. Variation in sequence of ی and ے

First Pass	Second Pass
FJL, FT, JUL, UL, NL	STCD, UHE, FA
بی	بی
بی بی	بے
بے	بیابان
بیابان	بی بی

2.1.6 ء (Hamza Character)

It is not clear that this character comes in the sorting sequence or not, e.g. whether ؤ is different from و for sorting purposes or not.

2.1.7 Harkaat (zer, zabar, pesh, khari zabar, tashdeed)

Normally zabar, zer, pesh respectively play a role in second pass of sorting, sequencing words with exactly same characters (e.g. بَن , بِن , بُن). However, it is not clear whether words without zer, zabar or pesh would come before or afterwards (e.g. what will be the sequence of words بَیل بیل).

Also, status of other aerab (diacritics) like Khari Zabar and tashdeed (e.g. sequence of the following words: پتا , پتّا , پتا) is not clear with reference to sorting sequence.

2.1.8 Miscellaneous

Status of Zer-e-izafat, Hamza-e-izafat and half space is not clear in sorting sequence. For example the following sequence of words can be sorted in different ways:

بانگ، بانگِ درا، بانگ دینا، بانگی

3. RECOMMENDED SOLUTIONS

In this section an effort is made to solve the collation sequence problems mentioned in the above section, solutions are recommended (as agreed by the panel of experts listed in Appendix A) and arguments are given to support them whenever required.

3.1 آ (Alif Mad)

There are three existing sequences in dictionaries as mentioned in section 2.1.1. The First question to be addressed is, whether آ should be considered a separate character or a variation of the already existing ا. If we go back to the origin of آ in Arabic, we find that آ is just a different stylistic way of writing two ا together. If we observe frequent use of آ we may argue that since آ is very productive as a starting sound in Urdu words, it should be considered as a separate character. But productivity of some sound does not necessitate that it should be considered as a separate character. We can see counter example of character ڑ where there is no word starting with it but it is considered as a character.

On the basis of above arguments آ should not be considered as separate character. Instead it should be considered as equal to ا following the Arabic tradition. Proposed sequence is as follows: (reading right to left)

ا، آ، آب، آپ، اب، ایوان

3.2 ں (Noon Ghunna)

There are two issues pertaining to ں. The first is whether it is an independent character or not. Observation of the use of ں in Urdu shows that it serves to add a "nasalization" feature to the preceding vowel rather than representing a separate sound. ن on the other hand is a nasal consonant itself. Therefore it is more appropriate to categorize ں as variation of ن instead of considering it an independent character.

The second issue is how they are placed in the collation sequence. Since ن is the independent character it was placed before ں. So proposed sequence is as follows: (reading right to left)

مان، ماں

3.3 ھ (Do Chashmey Hay)

Three possible sequences involving ھ are given in previous section 2.1.3, which are used in different dictionaries. Another sequence is possible where ھ and ہ be considered in the first pass as different characters and ہ comes before ھ. This will result in the following sequence: (reading right to left)

باپ، بہن، بہنگی، بھا بی، بھنگی، بیٹا

Regarding the problem that whether ھ is independent or a variation of ہ, the conclusion was that ھ should not be treated as an independent character because it does not add a new sound instead it adds the feature of aspiration to the preceding character, changing it to a new sound. The new sound is different from بہ (which is a sequence), thus بھ should form a new character.

There are some examples in other languages where combination of two characters changes to new sound but it is not considered in sorting sequence e.g. 'ch' in English but the two characters are considered individually in the sorting sequence. However a counter example of Spanish also exists where 'ch' is a character and sorts differently from 'c'.

Character combination with ھ should be considered as a separate character placed following all the words of that character, giving the following sequence: (reading right to left)

باپ، بہن، بہنگی، بیٹا، بھا بی، بھنگی

3.4 ۃ (Gol Tay)

The character ۃ is borrowed from Arabic. In Arabic *Tay* sound is represented by ت when it is part of the root morpheme. However, if the sound occurs as part of feminine or other suffix, *Tay* sound is represented as ۃ. ۃ is pronounced as ہ whenever there is a pause observed after the word ending with ۃ. However, this rule has not been observed in Urdu. But it is clear that ۃ is considered a variation ت as it represents the same sound *Tay*. Since ت is a regular character and ۃ is its variation, by following the rule applied in ن and ں case, the suggested sequence is as follows (going right to left):

ہ ... ۃ ت

3.5 ی/ے (Chhoti and Bari Yay)

It is conventional to put ی before ے. But the data collected shows that it is not decided that distinction between ی and ے should be in first pass or in second pass. Since ی and ے are two different separate characters there seems no reason to consider them in second pass. So proposed sequence is as follows:

بی، بی بی، بے، بیا بان

3.6 ء (Hamza Character)

As mentioned in section 2.1.6 the role of ء in the sorting sequence is ambiguous. ء seems to be a diacritic

mark due to its way of placement in words. It very often rests above some other character, making its appearance similar to aerab. But careful analysis of its placement shows that unlike aerab it requires a character indicator ("kursi") when used word medially, for example گئی.

There is an exception to this rule in the case of و, where ء is placed right above و without any character indicator [3] .

Once accepted as a character, the position of ء in collation sequence should be the traditional one i.e. between ہ and ی.

3.7 Harkaat (zer, zabar, pesh, khari zabar, tashdeed)

zabar, zer, pesh specify the content of words where as tashdeed and khari zabar change or add new content to existing phonemes. These two sets of aerab should be dealt with in different manners. Traditional sequence for zabar, zer, pesh is defined as stated. Word with no aerab should come at the end, after these diacritics.

Since tashdeed and khari zabar modify the existing character so following the previously defined rule in case of ن and ں, words without tashdeed should come before the one with tashdeed, after all other rules have determined the words to be equal. Word with khari zabar will come after same word without khari zabar. This is because khari zabar is a non-traditional way of writing Alif and thus comes after traditional way of writing same word without this Khari Zabar.

3.8 Miscellaneous

Zer-e-Izafat conveys linguistically different meaning but still it is same diacritical mark as zer. So for sorting purposes Zer-e-Izafat should be considered equivalent to zer. In case of compound words, space should be ignored. Sub entries are required to keep phrases with same first word together as followed by other languages.

4. CONCLUSION

Standardization of collation sequence for Urdu will make possible sorting the data and therefore proper use of spread sheets and databases for Urdu. Keeping in view its importance, this issue needs to be urgently resolved so that Urdu may be effectively used with computers for advanced data processing. Currently these recommendations have been forwarded to National Language Authority for ratification. Once finalized, they will be moved to international for a like ISO and Unicode for international standardization.

5. REFERENCES

[1] Sarmad Hussain, M. Afzal, *Urdu Computing Standards: Urdu Zabta Takhti (UZT) 1.01*, IEEE INMIC, Dec 2001

[2] M. Afzal, Sarmad Hussain, *Urdu Computing Standards: Development of UZT 1.01*, IEEE INMIC, Dec 2001

[3] قومی کونسل برائے فروغ اردو زبان،اردو املا،رشید حسن خان

۱۹۹۸،نئی دہلی،

[4] لاہور سنز، فیروز جامع، فیروز اللغات (FLJ)

[5] Standard Twentieth Century Dictionary: Urdu to English, Educational Publishing House, New Dehli, India (STCD)

[6] اسلام آباد، مقتدرہ قومی زبان ، فرہنگِ تلفظ (FT)

[7] اسلام آباد، مقتدرہ قومی زبان ، جدید اردو لغت (JUL)

[8] کراچی ، اردو لغت بورڈ، اردو لغت (UL)

[9] A Dictionary of Urdu, Classical Hindi and English, Crosby Lockwood and Son, London (1911) (UHE)

[10] دہلی ، فرہنگ آصفیہ (۱۹۱۸) (FA)

[11] لاہور ، سنگ میل، نور اللغات (NL)

APPENDIX A: ATTENDEES OF URDU COLLATION SEMINAR

1. Dr. Fateh Muhammad Malik, Chairman, National Language Authority, Islamabad (chair)
2. Parvez A. Butt, Project Director, TReMU, Islamabad
3. Mazhar Mahmood Shirani, Assoc. Professor (retd.), Govt. College, Sheikhupura
4. Suheyl Umar, Director, Iqbal Academy, Lahore
5. Muhammad Ahsan Khan
6. Liaqat Ali Asim, Editor, Urdu Dictionary Board, Karachi
7. Rafiuddin Hashmi, Prof. (retd.) Oriental College, Punjab Univ., Lahore
8. Mueen uddin Aqeel, Professor, Urdu Department, Karachi Univ., Karachi
9. Dr. Khurshid Rizvi, Professor (retd.), Arabic Department, Punjab Univ., Lahore
10. Zahid Munir Amir, Assistant Professor, Urdu Department, Punjab Univ., Lahore
11. Dr. Muhammad Afzal, Professor, Center for Information Technology, Arid Agriculture University, Rawalpindi
12. Dr. Qaiser Durrani, Professor, NUCES, Lahore
13. Belal Hashmi, Assoc. Professor, CRULP NUCES, Lahore
14. Shafiq ur Rahman, Assoc. Professor, CRULP NUCES, Lahore
15. Dr. Sarmad Hussain, Assoc. Professor, CRULP NUCES, Lahore
16. Nayyara Karamat, Research Officer, CRULP NUCES, Lahore
17. Tahira Naseem, Research Officer, CRULP NUCES, Lahore
18. Muhammad Usman Afzal, Research Officer, CRULP NUCES, Lahore

Advance Written Comments:

Dr. Khaver Zia, Professor, Beaconhouse Informatics, Lahore