Collation Sequence in Nepali

Srishtee Gurung and Laxmi Prasad Khatiwada
Madan Puraskar Pustakalaya
srishtee@mpp.org.np, laxmi@mpp.org.np

Abstract

This document approaches the research on collation sequence of Nepali Language through the scientific study of the language. It mainly focuses on the aspects of the writing system of the language that defines the collation for it. It argues for the new proposal of Nepali Collation sequence. It gives cases that dictate the collation or character ordering and invalidates the collation used till date.

1. Background

1.1. Languages in Nepal

The 2001 census has identified 91 languages spoken as mother tongues and other one unit of miscellaneous languages has been identified. Besides, a number of languages have been reported as 'unknown' languages (CBS, 2001), which need to be precisely identified on the basis of field observation and its analysis. Only a few of Nepal's languages have literate traditions. They include Nepali, Maithali, Tibetan/Sherpa, Newar, Limbu, Bhojpuri, Awadhi, and Lepcha in particular. These languages have employed various writing systems or scripts.

Table 1 at "appendix A" presents the distribution of the population of Nepal in terms of the mother tongue (1952/54 – 2001) showing the more significant languages divided into major language groups. The number of speakers and the percentage of population are based on the 2001 census. Many of these languages possess just a few thousand speakers or even less and the majority of these small population languages belong to the Tibeto-Burmese group. As per the statistics provided by Kathmandu Nepal, 2003 there are also some small number of speakers of the Austo-Asiatic language Satar and the Dravidian language Dhangar [2].

1.2. Writing systems of Nepal

The root of all writing systems of South Asia is the Brahmi system, which was created around 2,300 years ago. All of its derivatives and the Brahmi itself are alphabetic writing systems. The languages of South Asia and neighbouring areas have developed their own writing system, for instance, Tibetan, Nepali, Newari, Hindi, Tamil, Thai etc. Some of these scripts are so much similar as in the case of Devanagari system for Hindi with Newari and Nepali that one may find them very much alike. However, they are different both in terms of style and superficial appearance and very essence of writing, i.e. composition of writing [2].

2. Introduction

2.1. Nepali Language

Irrespective of its small size, an amazing cultural diversity that includes linguistic plurality is evident in Nepal. This multilingual setting confers on Nepal a distinctive position on the linguistic map of the world and renders it as one of the most fascinating areas of linguistic research [3].

In the case of Nepal, it is quite common to find people speaking different languages settled in a common locality. Hence, the need of a linking or communicating language for carrying out interpersonal and socio-economic activities. For this purpose, most non-Nepali speakers use Nepali as a lingua franca. Apart from the usage of Nepali as a lingua franca, Nepali is also used as a medium of instruction in education, administration, communication media etc. As a result , the use of two or more languages (bilingualism/ multilingualism) has been established well in the Nepal [3].

2.2. The Writing System of Nepali Language

This section describes the scientific method of writing Nepali language using Devanagri script. This script is not only used for writing Nepali but also for writing Indic Languages such as Hindi, Marathi Kashmiri and even Sanskrit. It is the descendant of Brahmi family which has been traced back to 500 B.C. B.C. The very word "Devanagari" originated from the Sanskrit words – Deva, meaning 'God' and Nagari, standing for 'city'. Thus , literally it means the "City of Gods". Devanagri is written from left to right and has no case distinction. The current writing system of Nepali Language has been revised recently [4].

Devanagri consists of 34 consonants (vyanjan), and 12 vowels (swar). In addition to it, there are other symbols such as diacritics and punctuation marks.

2.2.1. Nepali Vowels(swar)

The vowels are categorized in two classes:

- Independent vowels
- Dependent vowels

Independent vowels

अ आ इ ई उ ऊ ऋ ए ऐ ओ औ अं अः

are the independent vowels. They are given the first priority in the collation.

Dependent vowels

ा ि ी ु ू े ै ो ौ

Dependent vowels are those vowels which must exist with a consonant. Sometimes user becomes confuse due to the lack of analytical approach of Nepali character.

2.2.2. Nepali Consonants (Vyanjan)

क ख ग घ ङ च छ ज झ ञ ट ठ ड ढ ण त थ द
ध न प फ ब भ म य र ल व श ष स ह

Consonants in Devanagri have implied vowels where the character क actually means क and अ, ख means ख and अ and so on. Therefore removing the implicit vowel a gives either a dead consonant or a pure consonant.

For e.g. removing अ from क we get क्. Here क् is the pure consonant. When pure consonants come in contact with other consonants the pure consonants become dead consonants. For e.g. क् + ज=क्ज here the character 'half ka' is dead consonant as it is converted to its half form.

2.2.3. Diacritics symbols

ं ँ ः

These three characters are the diacritics marks used in Nepali Language.The diacritics can exist with both the consonants and independent vowels. For e.g.: आ + ँ + प = आँप, and अ + ं + श =अंश, स्वतः

2.2.4. Punctuation

् । ॥

There are only two additional punctuation marks in Devanagri apart from the punctuation marks used in English. ् । and .

- ् is used to reduced the property of an entity or a character. Like काम्द is equivalent to काम्द and गम्ला is equivalent to गम्ला.
- Known as Purnaviram is used for sentence break. राम घर गयो ।.
- is known as Double danda is a punctuation sign used to indicate the end of verse in a poem. However it is not used in Nepali.
- Punctuation mark '.' is used for allophonic modification like. However this punctuation is not used in Nepali Language though it exists in the Devanagri script.

2.3. Traditional writing system of Nepali

Traditional character set of Devanagri consists of:

2.3.1. Vowels

Independent vowels:

ॐ अ आ इ ई उ ऊ ऋ ऐ

ओ औ अं अः

Dependent vowels:

ँ ा ि ी ु ू े ै ो ौ

2.3.2. Consonants

क ख ग घ ङ च छ ज झ ञ ट ठ ड ढ ण त थ द ध न प फ ब भ म य र ल व श ष स ह क्ष त्र ज्ञ

The last three characters are treated as separate characters in this set. However they are treated as ligature in the current context of writing system.

2.3.3 Punctuation

् । ॥

Punctuation marks are used as it is in the current writing system for Nepali.

The collation sequence followed in traditional Nepali Language is as the following:

ॐ अ आ इ ई उ ऊ ऋ ए ऐ ओ औ अं अः क ख ग घ च छ ज झ ञ ट ठ ड ढ ण त थ द ध न प फ ब भ म य र ल व श ष स ह क्ष त्र ज्ञ ् । ॥

In this set the last three characters for consonants are treated as part of the character set. Hence they were put in the collation sequence for sorting order. However these three characters are ligatures. (Ligature will be discussed later in detail in section 5.1). Though the concept of ligature existed in the traditional Devanagri script but it was not applied to these three characters. This Traditional writing system is been followed in all the educational sectors such as schools and junior colleges from time immemorial. Unfortunately this incorrectness in character set was never acted upon and as an result we have been learning the incorrect collation sequence of Nepali Language. Further more the traditional collation sequence does not address the issue of other orthographic symbols such as 'Nukta' and the diacritic marks ं ँ ः. In traditional Devanagri script these character are treated as vowels. Although they can be combined with any other consonants or vowels they are not addressed as diacritic marks. In addition to these above problems, various other issues that were not addressed will be discussed later in the section 3.

The above problem arose mainly because there was a lack of scientific linguistic study of Nepali Language. The past trend for Nepali language targeted scientific studies on grammatical rules and spell checking for the language giving less priority to the most basic part of the script the character set and hence affecting the collation sequence.

According to the study of the linguistic aspect of the language it was found that a new proposal for the collation sequence had to be prepared. So a research work at Madan Puraskar Pustakalaya was carried on. Madan Puraskar Pustakalaya collaborated with Tribhuvan University for the scientific research of the language. Together with Ministry of Education, Curriculum development Center, Nepali Language in Information Technology (Steering Committee), Ministry of Environment and Science and Technology and Microsoft Nepali Localization Component, Madan Puraskar Pustakalaya has tabled a proposal for new character set and Collation sequence.

The proposed collation sequence is in accordance with Unicode encoding. It is given in the table format in "Table 2. The proposed Collation Sequence for Nepali Language in Appendix A".

3. Methods

The method that is used for the collation sequence is research oriented. A scientific study on linguistic area was conducted on the various factors of linguistic area that directed collation sequence for Nepali language. A group of linguist discussed each factor and passed the order of each character. They studied the current writing system of Nepali Language and compared with the actual scientific method of writing it. If difference is found the order of the character is determined and kept in the new proposal paper for collation sequence of the language.

4. Results

The comparative study between the traditional writing system and the scientific method of writing Nepali language using Devanagari script resulted in tabling new proposal for both character set and collation sequence for the language. The new collation sequence for the language is given in Appendix A in "Table 2 Proposed Collation Sequence for Nepali Language".

5. Discussion

The various factors governing the collation sequence to be prepared for Nepali Language are:

5.1. Ligature

Often consonants conjoin with one or more than one consonants thus suppressing the inherent vowel. The resulting form is termed as a ligature. In most cases, ligatures are the simple adjoinment of individual constants [4] .

E.g. घ्+ र= घ्र, द्+ व = द्व,श्+ र = श्र, ट् + र= ट्र

However, this may not be the case for some ligatures, where the adjoinment is not that straightforward and easily recognizable.

क्+ ष =क्ष, द् + य= द्य, त् + र= त्र, ज् +ञ= ज्ञ

These are known as compound phonemes. These are not independent single character so it is not necessary to include the whole character or phoneme in collation sequence.

क् + ष = क्ष, it comes after क and before ख

त् + र = त्र, it comes in between त and थ

ज् + ञ = ज्ञ it comes in between ज and झ respectively.

5.2. Diacritics

Some definitions of diacritics in the linguistic level:

5.2.1. ँ

Though ँ (आँप) is a nasalized sound and is used as a nasal vowel in the traditional Nepali language, in Unicode context ँ it is used as diacritic mark. Hence this is given the first priority in the collation sequence.

5.2.2. ं

(अंश) is a nasal vowel sound. In Unicode context ं is used as diacritics. In some cases ं is used as an equivalent of half म. For ex: अंबा and अम्बा (guava), संपादक and सम्पादक (editor) have the same meaning.

Most people use this symbol because if its easiness. In the traditional writing system it is the second last character in the vowel set whereas the scientific linguistic study revealed that it is the second character in the vowel set.

5.2.3. ः

(स्वत :) is an aspirant break_sound in Nepali and used as a diacritic mark. This is the sign of long stop. It is equivalent to the symbol half hah ie ह्. It does not exist in spoken Nepali such words are derived from Sanskrit. This is the last character in vowel set.

दु:ख(Grief), नि:शुक(free)

5.2.4. .

Is a glyph which is an alternative element of certain phoneme comes with an allophone such as ड़, ढ़, फ़

5.3. Punctuation

़ । ॥

These are the punctuation symbol which exists in Nepal. ़ । । Single danda or full stop. This is ending punctuation of a sentence. In other words। is used for sentence break. रामले भात खायो।

॥ Double danda is used in especially full stop for fictile. ॥ is used in Vedic Sanskrit and is not used in Nepali writing**.** So it is not prescribed for proposed Character consonant.

् is used to reduced the property of an entity or a character. Like काम्दा-कामदा टन्न-टन्न,

These are nasal sound rarely used in Nepal and are words borrowed form Tibeto-Burman. Since half form of alphabet and consonant written with this character have same weight in the collation, their order is determined by the immediate next character that follows it.

5.4. Abbreviation

This is like an abbreviation punctuation symbol and it comes after the numerals in the collation sequence.

हेर्नु=हे०, विशेषण=वि०, उपर्सग=उ०

5.5. Extra characters

There are some additional scripts which are used in oral Nepali. However these scripts may not be used exactly as they are in the writing system of the language. Such additional character include ऋ for vowels, ङ, ञ, ण for consonants',' '.' and '०' for punctuation and क्ष, त्र, ज्ञ for conjuncts. These additional scripts occur in Nepali Language as they are derived from Sanskrit. Therefore there is lot of words in Nepali written in such traditional script. Therefore Nepali Language has to accept these additional symbols.

ई ऊ ऋ ङ ञ ण ं ः ् ० श ष क्ष त्र ज्ञ ा ि ी ु ू े ै ो ौ

Their sequences are given in the proposal Table 1 in appendix A.

5.6. Position of Extra vowel in the collation sequence

ई ऊ ऋ अं अ

These are the extra vowel symbol in the sense that these vowels do not exist in oral Nepali but exists in writing system. These vowels make semantic contrast too.

ई = ई or इ are same in spoken system but are assumed as short and long vowel. i.e. ई = ii and इ = i in written Nepali.

5.7. Contrastive Analysis

ि and ी

जाति (Race/Caste) and

जाती (humble/polite)

ु and ू

जुन (relative pronoun) and जून(moon)

ऊ= ऊ or उ are same in spoken but assumed as short and long uu i.e. uu = ऊ and u = उ

ऋ This is the conjunct of R and I and etc but a specific identity in words like:

ऋषि	saint
ऋचा	hymn.stanza

5.8. Trends in writing system of Nepali:

Certain words in Nepali language are written using two different characters or ligatures. Such character are: ञ,न and ण. Replacing one character with the another do not alter the meaning of the word: for ex: When the character ञ in the word 'कञ्चन' is replaced with the character न making it as कन्चन, or मञ्जन as मन्जन the meaning of the word remains same. Even though scientifically क्ञचन is the correct way of writing, however writing कन्चन will have no effect on the collation sequence.

6. Conclusion

The problem existed in the traditional character set had to be dealt with. Addressing all the above problems, a new Character set need to be proposed.

Thus the collation sequence of the character had to be addressed in the new proposal too.

7. References

The following resources were consulted while preparing this document:

[1] Chapagain, N., 2053 B.S. Adhunik Nepali Vyakarna tatha rachana/ Ratna Pustak Bhandar Kathmandu, Nepal

[2] http://www.nepaldemocracy.org/media/Nepali%20Font%20Standards.pdf

[3] http://www.kantipuronline.com/pdf/tkp5.pdf

[4] http://en.wikipedia.org/wiki/Devanagari

[5] http://www.kellybadge.co.uk/MasterCatalogueDec04.pdf

[6] Yadav, Y., Regmi, B N. 2058 B.S. Bhasha Vigyan/New Hirabooks interprises, Kirtipur, Nepal

[7] Adhikari, Dr. H. 1998 A.D Samasamayik Nepali Vyakarna/Vidyarthi Pustak Bhandar, Kathmandu, Nepal

[8] Pokhrel, Dr. M.P. 2057 B.S. Dhwani Vigyan ra Nepali Bhashako Dhwani Parichaya/
Royal Nepal Academy, Kathmandu, Nepal

[9] Bhaskar, K. *Nepali Brihad Shabdakosh*, Royal Nepal Academy Kathmandu, 1984

[10] Macdonell, A. A. *A Vedic Grammar for Students,* Oxford University Press, USA; New Ed edition (November 1, 1976)

Appendix A

Table 1. Distribution of population of Nepal by Mother Tongue (1952/54-2001)

Table 4.1: Distribution of population of Nepal by mother tongue (1952/54-2001).

Mother Tongue	Population											
	1952/54	%	1961	%	1971	%	1981	%	1991	%	2001	%
A. Indo-European	6351899	77.13	7449604	79.14	9062435	78.42	12417886	82.66	14701283	79.50	17982769	79.1
1. Nepali	4013567	48.74	4796528	50.96	6060758	52.45	8767361	58.36	9302880	50.31	11053255	48.61
2. Maithili	1024780	12.44	1130402	12.01	1327242	11.49	1668309	11.11	2191900	11.85	2797582	12.30
3. Bhojpuri	477281	5.80	577357	6.13	806480	6.98	1142805	7.61	1379717	7.46	1712536	7.53
4. Tharu	359594	4.37	406907	4.32	495881	4.29	545685	3.63	993388	5.37	1331546	5.86
5. Awadhi	328408	3.99	477090	5.07	316950	2.74	234343	1.56	374635	2.03	560744	2.47
6. Rajbansi	35543	0.43	55803	0.59	55124	0.48	59383	0.40	85558	0.46	129829	0.57
7. Hindi	80181	0.97	2867	0.03	..	..	..	..	179997	0.92	105765	0.47
8. Urdu	32545	0.40	2650	0.03	..	..	..	..	202208	1.09	174840	0.77
B. Sino-Tibetan	1795337	21.08	1813083	19.26	1982635	17.16	1811944	12.06	3098698	16.76	4183995	18.4
9. Tamang	494745	6.01	528812	5.62	555056	4.80	522416	3.48	904456	4.89	1179145	5.19
10. Newar	383184	4.65	377721	4.01	454979	3.94	448746	2.99	690007	3.73	825458	3.63
11. Magar	273780	3.32	254675	2.71	288383	2.50	212681	1.42	430264	2.3	770116	3.39
12. Rai, Kirat	236049	2.87	239745	2.55	232264	2.01	221353	1.47	439312	2.38	..	..
13. Gurung	162192	1.97	157778	1.68	171609	1.49	174464	1.16	227918	1.23	338925	1.49
14. Limbu	145511	1.77	138705	1.47	170787	1.48	129234	0.86	254088	1.37	333633	1.47
15. Bhote, Sherpa	70132	0.85	84229	0.89	79218	0.69	73589	0.49	121819	0.66	129771	0.57
16. Sunuwar	17299	0.21	13362	0.14	20380	0.18	10650	0.07	..	..	26611	0.12
17. Danuwar	9138	0.11	11625	0.12	9959	0.09	13522	0.09	23721	0.13	31849	0.14
18. Thakali	3307	0.04	6432	0.07	..	..	5289	0.04	7113	0.04	6441	0.03
C. Austro-Asiatic	16751	0.20	29485	0.31	23853	0.21	28208	0.19	33332	0.18	40260	0.2
19. Satar	16751	0.20	18840	0.20	20660	0.18	22403	0.15	25302	0.14	..	..
20. Santhali	..	..	10645	0.11	3193	0.03	5804	0.04	8030	0.04	40260	0.18
D. Dravidian	..	..	..	..	..	..	..	..	15175	0.1	28615	0.1
E. Other	70340	0.85	114392	1.22	487060	4.21	764802	5.09	648627	3.51	28615	0.13
F. Not Sated/Unknown	752	0.01	6432	0.07	..	..	..	..	9157	0.05	503295	2.2
Total	8235079	100.00	9412996	100.00	11555983	100.00	15022839	100.00	18491097	100.00	22738934	100.00

Source: Population Censuses (1952/54-2001).

Table 2. Proposed Collation Sequence for Nepali Language

LETTER		COMMENT
Description	**Glyph**	
Vowels		
Nasalized Diacritic chandrabindu	◌ँ	Diacritic for vowel nasalization
Nasalized Diacritic anuswar	◌ं	Diacritic for additional nasalized sound
Visarga Diacritic	◌ः	Aspirant break
avagraha	ऽ	Required for Sanskrit, Maithili
vowel letter short A	अ	
vowel letter long AA	आ	
vowel letter short I	इ	
vowel letter long II	ई	
vowel letter long U	उ	
vowel letter short UU	ऊ	
vowel letter vocalic R	ऋ	
vowel letter short E	ए	

vowel letter diphthong EI	ऐ	
vowel letter short O	ओ	
vowel letter diphthong AU	औ	
Consonants		
varg 1, velar plosive consonant Ka	क	
varg 1, velar plosive consonant KHa	ख	
varg 1, velar plosive consonant Ga	ग	
varg 1, velar plosive consonant GHa	घ	
varg 1, velar nasal consonant NGa	ङ	
varg 2, palatal affricate consonant Ca	च	
varg 2, palatal affricate consonant CHa	छ	
varg 2, palatal affricate consonant Ja	ज	
varg 2, palatal affricate consonant JHa	झ	
varg 2, palatal nasal consonant NYa varg 3, Alveolar plosive (retroflex) consonant TTa	ञ ट	
varg 3, Alveolar plosive (retroflex) consonant TTHa	ठ	
varg 3, Alveolar plosive (retroflex) consonant DDa	ड	
varg 3, Alveolar plosive (retroflex) consonant DDHa	ढ	

varg 3, retroflex nasal consonant NNa	ण	
varg 4, dental plosive consonant Ta	त	
varg 4, dental plosive consonant THa	थ	
varg 4, dental plosive consonant Da	द	
varg 4, dental plosive consonant DHa	ध	
varg 4, dental nasal consonant Na	न	
varg 5, labial plosive consonant Pa	प	
varg 5, labial plosive consonant PHa	फ	
varg 5, labial plosive consonant Ba	ब	
varg 5, labial plosive consonant BHa	भ	
varg 5, labial nasal consonant Ma	म	
non-varg Approximant consonant Ya	य	
non-varg Approximant consonant Ra	र	
non-varg Lateral consonant La	ल	
non-varg Labio-velar approximant consonant Va	व	
non-varg Post-alveolar fricative consonant SHa	श	
non-varg retroflex fricative consonant SSa	ष	

non-varg Alveolar-fricative consonant Sa	स	
non-varg Glottal-fricative consonant Ha	ह	
Nukta (But it is not in Practice nowadays)	़	Commonly used modifier for allophonic variation and borrowed words
Halant	्	Punctuation to reduce the entity of a character
Full stop Or PurnavirAma	।	
Sacred Character	ॐ	It is used in Hindu hymns in the beginning step.
Numerals	0	
Numerals	१	
Numerals	२	
Numerals	३	
Numerals	४	
Numerals	५	
Numerals	६	
Numerals	७	
Numerals	८	
Numerals	९	
Numerals	.	

Source: References [2] and [5].

Table 3. - An example of sorting

Word	Romanized	Internal form	Sort order
पुरिया	(puriya)	प ु र ि य ा	2nd
पति	(pati)	प त ि	1st
प्रयत्न	(prayatna)	प ् र य त ् न	3rd