

# The Sinhala Collation Sequence and its Representation in UNICODE

By Weerasinghe A. R., Herath D. L., Gamage K.  
Language Technology Research Lab,  
University of Colombo School of Computing, Colombo, Sri Lanka.

## Abstract

The alphabet of a language is perhaps the first thing we learn as users. The alphabet of our mother tongue would be the first alphabet we ever learn. And yet, a closer look reveals that there is much about such an alphabet that we have not explicitly specified anywhere. The Sinhala alphabet order is a prime example. We use it, recite it and yet would be hard pressed to define it explicitly.

Sinhala is spoken in all parts of Sri Lanka except some districts in the north, east and centre by approximately 20 million people. It is spoken by an additional 30,000 (1993) people in Canada, Maldives, Singapore, Thailand and United Arab Emirates. Sinhala is classified as an Indo-European language and used as an official language.

The UNICODE Collation Algorithm (UCA) is an attempt to make explicit the collation sequence of any language expressed in the UNICODE (or any other) coding system. In order to express the Sinhala collation sequence (alphabetical order) using UCA, the authors undertook the task of identifying unresolved issues facing the unambiguous definition of the order. This paper first describes the issues identified through this study, suggesting alternate solutions and recommending one of them. Finally, it sets out the recommended collation sequence for Sinhala in the form of the UNICODE collation specification. The outcome of this process is a unique and unambiguous expression of the Sinhala collation sequence which could be tested using existing tools and software environments.

**Keywords:** UNICODE, Sinhala, Collation, UNICODE Collation Algorithm, Localisation, Internationalisation.

## 1. Background

THE COLLATION ORDER OF ANY LANGUAGE is one of the most important issues that has to be resolved urgently in the process of standardising such a language. Since languages have been used for centuries by humans without worrying about their irregularities, often their constructs are extra logical. The collation sequence of most languages faces this non-logical nature. Steps are being taken to avoid these ambiguities and irregularities and also to formalise the collation sequence as much as possible since it plays a major role in the process of standardising the languages. This is particularly so in the case of electronic texts of a language since the computer needs explicit ordering information in order to process such a language.

To illustrate simply, even in the apparently well understood case of the English Latin-1 character set, the font itself does not encode order. If it did, words beginning with 'Z' will precede those beginning with 'a' since all upper-case letters precede all lower-case letters in all encodings (including UNICODE) of English. The case for Sinhala is no exception. In fact, as will be clear in the ensuing discussion, the Sinhala collation sequence demands us to take some decisions thus far not explicitly made for the Sinhala language as a whole.

## 2. Introduction to Sinhala Alphabet

The Sinhala alphabet consists of characters which represent almost all the sounds that can occur in the language. On the other hand, it is phonetically over-specified in that there are

multiple characters to represent the same sound: for example ළ (dental la) & ල (Alveolar la), න (dental na) & ඹ (Alveolar na), ට (voiceless ta) & ධ (voiced ta).

The commonly accepted *Mixed Sinhala Alphabet* has a set of sixty characters. This set of characters can be classified into three categories, namely vowels, semi-consonants and consonants.

**Vowels:** There are 18 vowels in the Sinhala alphabet, අ, ආ, ඇ, ඇ,...ඹ , ඹ.

**Semi-consonants:** there are two characters which can occur only with a vowel: ෝ and ෞ.

**Consonants:** there are 20 consonants in the alphabet: ක, ඛ, ජ, ඣ, ඤ, ඦ, ට, ඨ, ඩ, ධ, ඳ, ඹ, ය, ර, ඼, ල, ඿.

In addition to the above characters there is another set of symbols called vowel-strokes or '*pilli*', to represent vowel sound when vowels are combined with consonants. For example: ක + ආ -> කා (k + aa) and ක + ෝආ -> කො (ka + o)

There is generally no disagreement regarding the order of characters within vowels, semi-consonants and consonants, except for 'ඥ' and 'ඨ'. The relative order of these character sets is also well defined, i.e. vowels are followed by semi-consonants which themselves are followed by consonants.

## 3. Methodology

The methodology adopted in this study is to first gather existing views and traditions of Sinhala language collation from scholars, observe the collation sequence adopted by the

major standard dictionaries and to research how collation sequence is determined at various state organisations in performing their regular tasks.

It is expected that this kind of study would reveal the main issues related to the collation order of Sinhala and how these issues are addressed by scholars, in dictionaries and by organisational practice.

#### 4. Issues identified in Collation Order

The following issues\* were identified at the beginning of the study. With these issues in mind, prominent dictionaries were searched, the views and opinions of scholars were obtained, and the procedures followed by state institutions and organisations were observed. The following listing identifies five major issues (first five with associated levels of importance) and three less critical ones which need to be resolved in order to proceed with the specification of an unambiguous collation sequence for Sinhala.

##### Issue # 1 [Level 2]

The positions of ‘*anusvara*’ and ‘*visarga*’ in the Sinhala collation sequence. While this is not really ambiguous as far as many dictionaries and linguists are concerned, its place at the beginning of the UNICODE code chart made it an issue to be resolved.

##### Issue # 2 [Level 3]

The position of the ‘*hal*’ sign (halant form) of a consonant in the sequence. Many alphabets of Sinhala do not explicitly specify the place of the ‘pure consonant’ form (the so-called ‘vowel removed form’) of Sinhala letters. As such, there is common confusion as to its rightful place in the alphabetical order. For digital representation, this becomes an important issue to be resolved.

##### Issue # 3 [Level 3]

The positions of words containing *yansaya*, *rakransaya* and *rephaya* when there are two or more alternative forms available for the same word. While in general there is agreement that these ‘short forms’ are exactly equivalent to their non-shortened forms, in a digital representation a decision has to be forced as to which of them precedes the other.

##### Issue # 4 [Level 2]

Miscellaneous issues such as the archaic way of writing words such as කපයාලය and the irregular forms කපර or කර and even කපර. Though rare, the exact function and position of such words in a sorted list of words needs to be explicitly given to facilitate digital processing.

##### Issue # 5 [Level 3]

Whether ‘ඥ’ is a ligature of ජ and ඥ or a single letter; and the right position of ‘ඥ’. This again is an issue raised by its existence as a separate code point in the UNICODE code chart.

##### Issue # 6 [Level 0]

The position of the letter ඹ. The Latin symbol ‘f’ was superimposed on the Sinhala letter m to produce the symbol ඹ, before the symbol ඹ was introduced into the alphabet. For this reason, and the phonetic closeness of the sounds represented by ඹ and ඹ, the letter ඹ has been popularly placed after the letter m in many contexts. On the other hand, the letter ‘ඹ’ being the newest letter of the Sinhala alphabet, is also placed at the end of the list of consonants in the alphabet.

##### Issue # 7 [Level 1]

There is a rule in the Sanskrit writing system that the consonant that comes after a *rephaya* is doubled, for example the words වාචිනා, මාගීගම. The reason for this appears to be to display other vowel modifiers clearly when they are used with a consonant that comes after the *rephaya*, for example as in කාමීකල, නාකීකික, ධමේමන්මාදය. The problem arising with this kind of phenomenon when sorting is whether their positions should be considered based on this doubled form, or on their corresponding simplest form (as කාමක, මාග, නාකක, ධමේන්මාදය in the examples above).

##### Issue # 8 [Level 1]

Finally, the sort order of words which consist of intra-word spaces, for example, the name ද සිල්වා. The issue here is whether to consider this as two words, to ignore the space and consider the string as a single word, or to consider the entire string including the space as the single full word.

While these issues were the ones pre-identified in the study, the availability of online tools for testing any suggested collation sequence expressed in accordance with the UNICODE Collation Algorithm specification, allowed us to look for any other issues which might be ‘thrown up’.

#### 5. Observation made in Dictionaries

For the purpose of this study some of the main and popular dictionaries were selected from among the various Sinhala dictionaries published. These dictionaries were selected after considering multiple attributes such as their quality, quantity in circulation, real usage and the perceived degree of authority of their compiler(s).

The following were the dictionaries selected for the present study:

- [1] *Sri Sumangala Shabdakoshaya* compiled by Ven. Velivitiye Soratha Thero
- [2] *Sinhala Shabdakoshaya* published by the Department of Cultural Affairs
- [3] *A Sinhalese – English Dictionary* compiled by Rev. Charles Carter
- [4] *Prayogika Shabdakoshaya* compiled by Dr. Harishchandra Wijetunge
- [5] *Sinhala Vishvakoshaya* published by the Department of Cultural Affairs

\* N.B : Issues # 6, 7, 8 were not taken into consideration in the first phase of the survey. The Levels indicate the perceived severity of the issue concerned, 3 being the most critical.

Issue	Sri Sumangala Shabdakosaya	Sinhala Shabdakosaya	Carter's Sinhala English Dictionary	Prayogika Shabdakosaya	Sinhala Vishvakoshaya
<b>The positions of the 'anusvara' and 'visarga'</b>	In all the dictionaries	the 'anusvara' and 'visarga' come at the end of the vowels.			
<b>The position of 'hal' sign</b>	The criteria followed is described in the Section 6	A letter with 'hal' sign comes after that letter's vowel derivatives. (The criteria followed in the dictionaries is described in the Section 6)			
<b>The positions of the words containing 'yansaya', 'rakaransaya' and 'repaya'...</b>	Observations could not be made regarding this issue since no dictionary uses two or more forms of the same word.				
<b>Irregular forms</b>	ඹ is used only to represent කර not කර්	ඹ=කර් කර් is written as it is	ඹ is used only to represent කර not කර්	Both forms are followed in different places ඹ=කර් or කර්	
<b>Status of 'Z'</b>	-considered as a ligature-	-considered as a ligature-	-considered as a ligature-	-considered as a ligature-	-considered as different char-
<b>The letter *</b>	-not present-	* comes at the end of consonants	-not observed-	* comes at the end of consonants	* comes at the end of consonants
<b>Doubled Reph</b>	-given along with the main entry as alternative spelling-	-given along with the main entry as alternative spelling-	-given along with the main entry as alternative spelling-	-given along with the main entry as alternative spelling-	-not observed-
<b>Intra-word space</b>	-space has been considered-	-space has been considered-	-space has been considered-	-space has been considered-	-space has been considered-

Table 1: The position taken by dictionaries on the eight issues under consideration

The issues identified regarding the collation order of Sinhala were kept in mind while these dictionaries were being studied. The information gathered from these dictionaries regarding each issue is summarised in Table 1.

Based on the above, some partial conclusions could be made as follows:

**Issue 1:** This appears to be a non-issue as far as dictionary compilers are concerned. These two semi-consonants are placed at the end of the set of vowels in the Sinhala alphabet.

**Issue 2:** Apart from the *Sri Sumangala* dictionary, each of the other four had a clear decision that the 'hal' form comes after all other vowel derivatives. While the justification given by each of these two schools makes sense in their own contexts, the majority decision may need to be adopted for our purposes. Section 6 illustrates the difference between the two schemes.

**Issue 3:** Interestingly, none of the dictionaries are able to shed light on this issue owing to each only containing a single form – either the short or the non-short.

As such this issue cannot be resolved using this methodology.

**Issue 4:** There is wide variation on the treatment of this issue. As such, a final decision on resolving this is deferred at this stage.

**Issue 5:** All dictionaries implicitly consider 'ඹ' as a ligature by their positioning of words beginning with it appearing soon after those beginning with 'ඵ' and 'ඹ'.

**Issue 6:** In all dictionaries which includes it, the position of the letter ඹ is immediately after the consonants.

**Issue 7:** The doubling of the *reph*-modified consonant is given as a spelling variant of the simpler form in all dictionaries which contained it.

**Issue 8:** All dictionaries include words which have intra-word spaces where appropriate.

## 6. Procedures Followed by Dictionaries to Sort Words

The procedure followed in the *Sri Sumangala Sabdhakoshaya* to arrange words manually according to the alphabetical order is best specified by the following algorithm:

1. Identify the syllabic units\* of the two words  
Let the two words be  $w_1$  and  $w_2$
2. Write each syllabic unit of both words as a consonant-vowel pair†
3.  $i=0$
4. Choose the  $i^{th}$  character of each word  
Let the two characters be  $w_1$  (ch (i)) and  $w_2$  (ch (i))
- 4.1. If  $w_1$  (ch (i)) =  $w_2$  (ch (i))
  - 4.1.1.  $i=i+1$
  - 4.1.2. go to 4
- 4.2. Else if  $w_1$  (ch (i)) >  $w_2$  (ch (i))
  - 4.2.1.  $w_1 > w_2$
  - 4.2.2. break
- 4.3. Else
  - 4.3.1.  $w_1 < w_2$
  - 4.3.2. break

Dictionaries other than *Sri Sumangala Sabdhakoshaya* compare consonant-vowel pairs in a different manner which makes the two approaches different. In this method when two consonant-vowel pairs are compared two consonants and two vowels are compared separately. In the cases where vowels are

not present the consonant of the next consonant-vowel pair is not taken as in the *Sri Sumangala Sabdhakoshaya*.

**7. Views of Scholars/Academics and Linguists**

The following scholars and academics were consulted with a view to acquiring their expert views – often based on their respective linguistic persuasions. The aim of the consultation was to attempt to achieve consensus and not just for documenting their independent views.

- Professor Vinee Vitharana (VTH)  
*Chief Editor of the Sinhala Dictionary. Former Professor of Sinhala at the University of Ruhuna.*
- Professor Wimal G. Balagalle (WBA)  
*Former Chief Editor of the Sinhala Dictionary. Emeritus Professor of the University of Sri Jayewardenapura.*
- Professor W. S. Karunathilake (WSK)  
*Former Professor of Linguistics at the University of Kelaniya*
- Professor J.B. Dissanyaka (JBD)  
*Emeritus Professor of the University of Colombo*

Issues	VTH	WBA	WSK	JBD	SGA	HWI	RUP
1	At the end of the vowels						
2	At the end of the vowels						
3	Need to have a policy: what is simple should come first	Need to have a policy: whatever non-confusing	Need to have a policy	Need to have a policy: what is simple should come first	Need to have a policy: what is non-confusing	Need to have a policy: priority must be given to what is commonly written	Need to have a policy: priority should be given to the tradition written
4	ඉ=කරු කරු should be written as it is	ඉ=කරු කරු should be written as it is	Sanskrit loan words should be written in their traditional forms, but English loan words can be written in either form	ඉ= කරු කෘ= කරු	ඉ=කරු කරු should be written as it is	ඉ=කරු ඉ= කරු	ඉ=කරු කරු should be written as it is
5	'ඳ' is a ligature ජ+ ඳ	'ඳ' is a ligature ජ+ ඳ	'ඳ' is a ligature ජ+ ඳ	'ඳ' is a ligature ජ+ ඳ	'ඳ' is a ligature ජ+ ඳ	'ඳ' is a ligature ජ+ ඳ	'ඳ' is a ligature ජ+ ඳ
6	ඟ comes at the end of consonants	ඟ comes at the end of consonants	ඟ comes at the end of consonants	ඟ comes at the end of consonants	ඟ comes at the end of consonants	ඟ comes at the end of consonants	ඟ comes at the end of consonants
7	-give along with the simplest form-	-give along with the simplest form-	-give along with the simplest form-	-give along with the simplest form-	-give along with the simplest form-	-give along with the simplest form-	-give along with the simplest form-
8	-space has to be considered-	-space has to be considered-	-space has to be considered-	-space has to be considered-	-space has to be considered-	-space has to be considered-	-space has to be considered-

Table 2: The position taken by Linguists on the eight issues under consideration

\* Syllabic unit means entity that contains a consonant and a vowel which is represented with pilli, e.g. බි,කො, ක are syllabic units and in some contexts ක, අ,ආ can also be syllabic units, in අකො අ and ක are syllabic units. The syllabic units of අච්චකො are අ, ච, චි, ච්, ක † If the syllabic unit does not consist of a vowel write 'null' in place of vowel (e.g. ක = ක null). The 'null' is considered as a character and it is the character that has the greatest value in the weight space.

- Professor Sucharitha Gamlath (SGA)  
*Former Professor of Sinhala at the University of Ruhuna.*
- Dr. Harishchandra Wijethunge (HWI)  
*Author of the Prayogika Sinhala Sabdhakoshaya*
- Mr. Rupasinghe Perera (RUP)  
*Deputy Director, Pirivena Education Branch, Ministry of Education Secretary, Sri Lanka Oriental Languages Society*

Having explained the aims and objectives of this study, a list of lexemes that concretely represents all the possible issues was carefully designed and given to each consultant – rather than posing the issues in their abstract form. This approach forced an explicit response rather than inviting rigorous expositions of the theoretical basis for same. There were some issues which some of the linguists could not provide a direct answer to. However, most were able to make their suggestions as to how to resolve such issues by relying on their own linguistic theories. The books written by some of these scholars were also considered during this study. The comments made by each expert regarding the identified issues and their suggestions are summarised in Table 2.

Issues 1, 2, 6, 7 and 8 were not disputed by any of the experts who agreed with the majority (4) of mainstream dictionaries. While Issue 3 had no consensus solution, all experts agreed that there should be a single well-specified standard. Issue 4 too had no clear consensus except for the recommendation that ‘ඛඛ’ should be written as it is. There also appears to be a majority view that ‘Z’ should be treated as a ligature — with the only dissenting scholar too late arriving at a consensus in the interest of arriving at an overall consensus.

## 8. Procedures followed in State Institutions and Organisations

The following government organisations and institutes were selected for the purpose of identifying the different collation orders adopted by them for their regular work.

- **National Library & Documentation Centre (NLDC)**
  - a. An explicit alphabetical order is available at NLDC
  - b. The Sri Lanka National Bibliography is prepared according to this alphabetical order
- **National Institute of Education (NIE)**
  - a. The NIE has adopted the alphabetical order given in the *Sri Sumangala Shabdakoshaya*.
  - b. This order is followed when school text books and recommended books for school children are prepared.
  - c. The specified alphabetical order for government examinations (e.g.: GCE (O/L) and GCE (A/L)) is also the same.
  - d. Further recommendations of the NIE are given in a separate publication entitled *Sinhala Lekhana Reethiya*.
- **Public Library – Colombo (PUB)**
  - a. The alphabetical order given in the *Sinhala Encyclopedia* is followed.
- **Sinhala Dictionary Office (SDO)**
  - a. The criteria followed by the SDO is the criteria followed in the *Sinhala Sabdhakoshaya*
- **Sinhala Encyclopedia Office (SEO)**
  - a. – not yet responded –
- **Election Commissioner’s Office (ELE)**
  - a. An explicit alphabetical order is available at ELE
  - b. ‘anusvara’ and ‘visargaya’ comes at the end of vowels
  - c. ‘hal’ sign comes at the beginning of vowels
  - d. When there are two or more alternative forms available, the collation order is found according to the simplest form and the priority is given to the simplest form (issue #3)
  - e. The letter ‘Z’ is considered as the conjunction of ජ and ඛ.
  - f. The recommendations of the NIE given in the *Sinhala Lekhana Reethiya* are also followed by the ELE.
- **Library – University of Colombo (UOC)**
  - a. The alphabetical order given in the *Sinhala Sabdakoshaya* is followed.
- **Library – University of Peradeniya (PDN)**
  - a. The alphabetical order is the same as that used by UOC.
- **Library – University of Kelaniya (KLN)**
  - a. – not yet responded –
- **Library – University of Sri Jayewardenepura (SJP)**
  - a. –not yet responded –
- **Library – University of Ruhuna (RHU)**
  - a. The alphabetical order given in the *Sinhala Encyclopedia* is followed.

Of the above, the NIE and Election Commissioner's Office (ELE) deserve special attention. The ELE standard is of interest to this study because it explicitly addresses the issues at hand – Issues 1, 2, 3 and 5. In Issues 1 and 2, the ELE standard tallies with those of the majority of dictionaries and linguists. Interestingly, ELE has a definite recommendation for Issue 3, i.e. to locate all such form variations together at the rightful place of the simplest form with the simplest form preceding the other forms in decreasing order of simplicity. Finally, on Issue 5 (6, 7 and 8 also), the ELE standard concurs with that of the expert consensus.

Since the government recognises the NIE as the prime authority in setting educational standards the order recommended by them becomes of utmost importance. Some of the other reasons for attaching such importance to this recommendation include:

- The standard specified has been created by representative groups of scholars and linguists including many of those consulted in the present study.
- Generations of school teachers and students have already adopted this standard and hence it is the closest to a *defacto* standard.
- Their more recent publication, *Sinhala Lekhana Reethiya*, is widely used by state organisations including the Commissioner of Elections.

## 9. Summary Recommendations

The status of each of the issues considered in this study together with the recommended solution is presented below.

- Issue 1: The dictionary survey and ratified by the expert consultation resolved this issue to the satisfaction of the authors: treat both the '*anusvara*' and '*visarga*' as appearing in the alphabetical order immediately after all the vowels. This is also further confirmed by the ELE and NIE standards which are in wide practical use.
- Issue 2: The dictionary disparity with regard to the correct position for the 'hal' form was resolved by the unanimous opinion of the experts consulted that it should immediately follow the vowels but precede the '*anusvara*' and '*visarga*'.
- Issue 3: This was one of the issues on which empirical evidence was scarce. However, the openness of all the linguists for some standard and the simplicity rule recommended by some of them and clearly enshrined in the ELE standard is to order all forms of such words adjacent to each other beginning with the simplest form and increasing in complexity. This would prescribe the following order on the three common forms of the work *karyalaya*: කාර්යාලය, කාර්යාලය, කාර්යාලය.
- Issue 4: This is the issue with the greatest degree of divergence in opinion. Three of the dictionaries and five of the linguists however concurred that 'ඹ' is used only to represent 'ඹරු' not 'ඹරු'. The latter is repre-

sented as it is. This is in contrast to the original Sinhala UNICODE recommendation where 'ඹ' = 'ඹරු' and 'ඹ' = 'ඹරු'. It seems prudent to adopt the majority opinion.

- Issue 5: This was the single main success in the consensus-seeking process. It is thus recommended that 'ඹ' be treated as the ligature of 'ඹ' + 'ඹ' so that it does not appear in the order thought to be implied in the UNICODE code chart.
- Issue 6: This seemed to be an issue as it was introduced later to the Sinhala alphabet and the phonetic similarity of the letters 'ඹ' and 'ඹ'. There was confusion with the symbol 'ඹ' too. According to Sinhala Lekhana Reethiya, the book published by the NIE for Sinhala, and all the scholars it is accepted that the letter 'ඹ' should come at the end of the consonants.
- Issue 7: This form is used merely for representation purposes. The underlying meaning of both 'කාමිමක' and 'කාරුමක' is the same. Therefore they occupy the same collation position. In dictionaries these are given along with the main entry as spelling variations.
- Issue 8: It is important to consider intra-word space when sorting is done in some domains (e.g. directories of names). However, this cannot be prescribed in the alphabet — it can only be done at the level of the particular application.

Based on the above recommendations and extensive testing done using early versions of the proposed collation sequence, a UNICODE Collation Element table together with its weights is recommended as the explicit specification of the Sinhala alphabet for use in electronic processing of Sinhala. The documents observed at the organisation mentioned above and other relevant documents including the proposed Collation Element Table can be found at the URL <http://www.ucsc.cmb.ac.lk/ltrl/public/collationDocs.html>.

---

**The proposed Collation Element Table can be found at the URL <http://www.ucsc.cmb.ac.lk/ltrl/public/collationDocs.html>**

---

## 10. Conclusion

At the outset we pointed out that a complete and unambiguous specification of the Sinhala alphabet is an essential and urgent requirement for all kinds of electronic processing of Sinhala text. The process of study revealed five major areas unresolved as far as the Sinhala collation order was concerned and three other areas which needed clarification. We outlined a methodology of arriving at a set of well informed recommendations based on three sources: widely accepted dictionaries, the most respected Sinhala scholars and the most widely adopted official standards on collation sequence.

Using a consensus-based approach, we have successfully arrived at a unique collation sequence for the Sinhala language and expressed it explicitly using the UNICODE Collation Algorithm specification of the UNICODE Consortium. Testing of this specification for arbitrary lists of words is made possible by online tools available from International Components for UNICODE (ICU).

### Acknowledgements

This work was partially supported by IDRC under the PAN Localization Project and sponsored by the ICT Agency of Sri Lanka. The authors are indebted to all Sinhala Language scholars, Publishers and other Practitioners who collaborated willingly in the work described in this study. In particular we wish to acknowledge the contribution of Mr. Harsha Wijayawardhana and Mr. Asanka Wasala as well as other colleagues in the Language Technology Research Centre of the University of Colombo School of Computing.

### References

- [1] Karunathilake, W.S , Sinhala Basha Viyakaranaya, M.D. Gunasena & Company Ltd, 1997.
- [2] Wijetunge, H , Sinhala Akuru Akaradhie Kireema Pramitha Kireema, S. Godage Brothers, 2003
- [3] Perera, R, Sinhala Vahara Athpotha, Thivira Publishers, 2004
- [4] Sinhala Lekhana Reethiya, National Institute of Education, 2001
- [5] Disanayaka, J.B, Akuru ha Pilli, S. Godage Brothers, 2000



A. Ruwan Weerasinghe

A. Ruwan Weerasinghe obtained a first degree in Mathematics and Statistics from the University of Colombo, a Masters in Computer Science in 1990 and a PhD in Natural Language Processing in 1994 (both from the University of Cardiff, UK). He then returned to the University of Colombo and served as a Senior Lecturer — first at the Department of Computer Science and then the University of Colombo School of

Computing (UCSC). After being appointed Head of the Department of Intelligent Systems in 2003, he became Director of the School in May 2004. Ruwan may be contacted at [aru@ucsc.cmb.ac.lk](mailto:aru@ucsc.cmb.ac.lk)



Dulip L. Herath

Dulip L. Herath is a Research Assistant at the Language Technology Research Laboratory (LTR) at the University of Colombo School of Computing, Sri Lanka. He obtained his Bachelor of Science degree in Computer Science in 2004 from the University of Colombo and is currently reading for his MPhil in Natural Language Processing. His research interests are Natural Language Parsing, Localisation and Local

Language Processing. As a member of the LTR he is involved in localisation activities, namely defining the collation order of Sinhala, Interface Terminology Translation, and local language web content development using Unicode. Dulip may be contacted at [dherath@webmail.cmb.ac.lk](mailto:dherath@webmail.cmb.ac.lk)



Kumudu Nuyanie Gamage

Kumudu Nuyanie Gamage obtained a B.A. (Hons) in Linguistics in 2004 from the University of Kelaniya, Sri Lanka. She joined the Language Technology Research Laboratory (LTR) of the University of Colombo School of Computing as a Research Assistant the same year. As a member of a development team, Kumudu was involved in the theoretical side of Computational Linguistics with special reference to the Sinhala language. Presently she is working at the Department of Linguistics of

the University of Kelaniya, as an Assistant Lecturer. Her Research interests are in Computational Linguistics, including Syntax and Sinhala Grammar, and the Analysis of differences between Sinhala and Japanese Languages with a special focus on Sinhala-Japanese/Japanese-Sinhala Translation. Kumudu may be contacted at [kgamage@webmail.cmb.ac.lk](mailto:kgamage@webmail.cmb.ac.lk)