L2/06-105

INTERNATIONAL STANDARD

ISO/IEC 14651

First edition
2001-02-15

FDAM 3
2006-03-23

Information technology — International string ordering and comparison — Method for comparing character strings and description of the common template tailorable ordering

AMENDMENT 3

Technologies de l'information — Classement international et comparaison de chaînes de caractères — Méthode de comparaison de chaînes de caractères et description du modèle commun et adaptable d'ordre de classement

AMENDEMENT 3

Reference number
ISO/IEC 14651:2001/Amd.1:2006(E)

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Amendment 3 to ISO/IEC 14651:2001 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 2, *Coded Character Sets*.

Information technology — International string ordering and comparison — Method for comparing character strings and description of the common template tailorable ordering

AMENDMENT 3

Clause 3

Replace the normative references with the following.

ISO/IEC 10646:2003, *Information technology — Universal Multiple-Octet Coded Character Set (UCS)*

ISO/IEC 10646:2003/Amd.1:2005 *Information technology – Universal Multiple-Octet Coded Character Set (UCS: Architecture and Basic Multilingual Plane Amendment 1: Glagolitic, Coptic, Georgian and other characters.*

Subclause 6.5

Replace Subclause 6.5 with the following.

6.5 Name of the Common Template Table and name declaration

Whenever the Common Template Table is referred externally as a base point in a given context, whether in a process, contract, or procurement requirement, it shall be referenced using the name ISO14651_2005_TABLE1. If another name is used due to practical constraints, a declaration of conformance shall indicate how the correspondence between this other name and the name ISO14651_2005_TABLE1 is taken care of.

The use of a defined name is necessary to manage the different stages of development of this table. This follows from the nature of the reference character repertoire, for which development will be ongoing for a number of years or even decades.

Replace Annex A with the following.

Annex A
(normative)

Common Template Table

In order to minimize formatting problems and the risk of errors in reproduction, the common template table is provided separately in a machine-readable file as a normative component of this International Standard. The file name for this language version is different from the normative reference name specified in clause 6.5 of this International Standard due to the existence of file versions commented in other natural languages. The file for this language version can also be retrieved on the ITTF web site at the following URL:

ISO14651_2005_TABLE1_en.txt [final URL to be provided by ITTF at publication stage]

There is an official French version of the file which only differs in its comments (its technical content is identical), and its name is: ISO14651_2005_TABLE1_fr.txt

NOTE 1 This amendment deprecates, but does not preclude specific reference to, the previous tables, which contained and still contains respectively ordering information and. The previous tables can be found at the following URLs:

[ordering information on the repertoire of characters as defined in ISO/IEC 10646-1:1993 including Amendments 1-9] http://www.iso.org/ittf/ISO14651_2000_TABLE1.htm

[ordering information on the combined repertoire of characters of ISO/IEC 10646-1:2000 and ISO/IEC 10646-2:2001] http://www.iso.org/ittf/ISO14651_2002_TABLE1_en.txt

[ordering information on the repertoire of characters as defined in ISO/IEC 10646:2003]
http://www.iso.org/ittf/ISO14651_2003_TABLE1_en.txt

The current Common Template Table reflects the repertoire of characters as defined in ISO/IEC 10646:2003 including its amendment 1.

NOTE 2 The repertoire targeted by this amendment 3 to ISO/IEC 14651:2001 is equivalent to the repertoire of *The Unicode Standard Version 4.1, published by The Unicode Consortium*.

When ordering data applicable to other amendments of ISO/IEC 10646:2003 becomes available, this International Standard and specifically its Common Template Table will be amended accordingly to cover the ordering of the additional characters and scripts. To meet cultural requirements of specific communities, delta declarations will have to be applied to the amended table as defined in this International Standard.

ISO_14651_2005_TABLE1 is the name that is used for referring to this table in this version of this International Standard.

Include the following new section at the end of annex B:

B.5. Example 5 – A tailoring for Khmer

The Khmer script is mainly used in Cambodia. The tailoring given below is not included in the CTT (see annex A) itself in order to keep the CTT simple, especially for rare letterforms. E.g. the Khmer ROBAT for which the tailoring below may not be desirable for efficiency reasons, since this letter occurs very rarely, but the tailoring for handling it correctly may affect the efficiency of collation also for texts that do not contain any ROBAT.

```
reorder-after <MAX>


% Khmer:

collating-symbol <S1794_S17C9> % KHMER LETTER BA, KHMER SIGN MUUSIKATOAN

collating-symbol <S1794_S17CA> % KHMER LETTER BA, KHMER SIGN TRIISAP

collating-symbol <S17BB_S17C6> % KHMER VOWEL SIGN U, KHMER SIGN NIKAHIT

collating-symbol <S17B6_S17C6> % KHMER VOWEL SIGN AA, KHMER SIGN NIKAHIT

collating-symbol <C1780>..<C179C>


   % Declaration of Khmer contractions

collating-element <U1794_17C9> from "<U1794><U17C9>" % KHMER LETTER BA, KHMER SIGN MUUSIKATOAN

collating-element <U1794_17CA> from "<U1794><U17CA>" % KHMER LETTER BA, KHMER SIGN TRIISAP

collating-element <SW_17CC_1780>..<SW_17CC_17A2> from "<U1780>..<U17A2><U17CC>"

% KHMER LETTER KA, KHMER SIGN ROBAT..KHMER LETTER QA, KHMER SIGN ROBAT

collating-element <SW_17CC_17A5>..<SW_17CC_17B3> from "<U17A5>..<U17B3><U17CC>"

% KHMER INDEPENDENT VOWEL QI, KHMER SIGN ROBAT..KHMER INDEPENDENT VOWEL QAU , KHMER SIGN ROBAT

collating-element <U17C6_17BB> from "<U17BB><U17C6>" % KHMER VOWEL SIGN U, KHMER SIGN NIKAHIT (OM properly spelled)

collating-element <U17BB_17C6> from "<U17C6><U17BB>" % KHMER SIGN NIKAHIT, KHMER VOWEL SIGN U (OM with the wrong sequence of the characters)

collating-element <U17C6_17B6> from "<U17B6><U17C6>" % KHMER VOWEL SIGN AA, KHMER SIGN NIKAHIT (AM properly spelled)

collating-element <U17B6_17C6> from "<U17C6><U17B6>" % KHMER SIGN NIKAHIT, KHMER VOWEL SIGN AA (AM with the wrong sequence of the characters)

collating-element <U17D2_1780>..<U17D2_179C> from "<U17D2><U1780>..<U179C>"
```

Deleted: , Thai, and Lao

Deleted: , Thai, and Lao

Deleted: s

Deleted: are

Deleted: neighbouring countries, and the tailorings for them do not conflict in any way since they are three separate scripts. Therefore these three scripts should be covered by a single tailoring rather than one tailoring per script.

Deleted: for two reasons. Firstly,

Deleted: Secondly, the “swapping” for Thai and Lao “pre-vowels” is done by other means in the parallel standard UAX 10. But the base collation weighting table for UAX 10, the DUCET, is fundamentally (but not syntactically) basically the same as the CTT. Hence a Khmer/Thai/Lao tailoring of the DUCET would not include the swapping of “pre-vowels”. For a description of Thai (and to some extent Lao) collation requirements, see annex C.2 below

Deleted: % Thai:¶
collating-element
<U0E5A_0E30> from
"<U0E5A><U0E30>" % special
end symbol¶
% Thai pre-vowels:¶
collating-element
<U0E40_0E01>..<U0E40_0E2E>
from
"<U0E40><U0E01>..<U0E2E>"¶
collating-element
<U0E40_0E40_0E01>..<U0E40_0E
40_0E2E> from
"<U0E40><U0E40><U0E01>..<U0E
2E>"¶
collating-element
<U0E41_0E01>..<U0E41_0E2E>
from
"<U0E41><U0E01>..<U0E2E>"¶
collating-element
<U0E42_0E01>..<U0E42_0E2E>
from
"<U0E42><U0E01>..<U0E2E>"¶
collating-element
<U0E43_0E01>..<U0E43_0E2E>
from
"<U0E43><U0E01>..<U0E2E>"¶
collating-element
<U0E44_0E01>..<U0E44_0E2E>
from
"<U0E44><U0E01>..<U0E2E>"¶
¶
collating-element
<U0E4D_0E32> from
"<S0E4D><S0E32>" % AM (in
the wrong order really)¶
¶
% Lao:¶
% Lao pre-vowels:¶
collating-element
<U0EC0_0E81>..<U0EC0_0EAE>
from
"<U0EC0><U0E81>..<U0EAE>"¶
collating-element
<U0EC0_0EC0_0E81>..<U0EC0_0E
C0_0EAE> from ... [1]

```
% COENG, KHMER LETTER KA..COENG, KHMER LETTER QA

collating-element <U17D2_17A5>..<U17D2_17B3> from "<U17D2><U17A5>..<U17B3>"

% COENG, KHMER INDEPENDENT VOWEL QI..COENG, KHMER INDEPENDENT VOWEL QAU


reorder-after <S1794> % KHMER LETTER BA

<S1794_17C9> % KHMER LETTER BA, KHMER SIGN MUUSIKATOAN

<S1794_17CA> % KHMER LETTER BA, KHMER SIGN TRIISAP
```

Deleted: reorder-after <S0ECD> % LAO NIGGAHITA¶
<S0EB3> % LAO VOWEL SIGN AM¶

```
reorder-after <S17C5> KHMER VOWEL SIGN AU

<S17BB_17C6> % KHMER VOWEL SIGN U, KHMER SIGN NIKAHIT


reorder-after <S17C6> KHMER SIGN NIKAHIT

<S17B6_17C6> % KHMER VOWEL SIGN AA, KHMER SIGN NIKAHIT


reorder-after <S17D2>

<C1780>..<C1794> % COENG, KHMER LETTER KA..COENG, KHMER LETTER BA

<C1795>..<C179A> % COENG, KHMER LETTER PHA..COENG, KHMER LETTER RO

<C17AB> % COENG, KHMER INDEPENDENT VOWEL RY

<C17AC> % COENG, KHMER INDEPENDENT VOWEL RYY

<C179B> % COENG, KHMER LETTER LO

<C17AD> % COENG, KHMER INDEPENDENT VOWEL LY

<C17AE> % COENG, KHMER INDEPENDENT VOWEL LYY

<C179C>..<C17A2> % COENG, KHMER LETTER VO..COENG, KHMER LETTER QA


reorder-after <SFFFF>

order_start forward;forward;forward;forward


<U1794_17C9> <S1794_17C9>;<BASE>;<MIN>;<U1794_17C9> % KHMER LETTER BA, KHMER SIGN MUUSIKATOAN
```

Deleted: <U0E5A_0E30> IGNORE;IGNORE;IGNORE;<U0E5A_0E30> % THAI CHARACTER ANGKHANKHU, THAI CHARACTER SARA A¶

```
<U1794_17CA> <S1794_17CA>;<BASE>;<MIN>;<U1794_17CA> % KHMER LETTER BA, KHMER SIGN
TRIISAP

%% The ROBAT contractions should be used only in an "advanced" tailoring for

%% Khmer, since ROBAT is rather rarely used, and these contractions

%% may impact on the efficiency of the key computation even if ROBAT does not

%% occur, since these contractions begin with commonly used letters.

<SW_17CC_1780>..<SW_17CC_17A2>                  "<S179A><S17D2><S1780>..<S17A2>";
"<BASE><VRNT1><BASE><BASE>";"<MIN><MIN><MIN><MIN>";
<SW_17CC_1780>..<SW_17CC_17A2>

% KHMER LETTER KA, KHMER SIGN ROBAT..KHMER LETTER QA, KHMER SIGN ROBAT

<SW_17CC_17A5>..<SW_17CC_17A6>           "<S179A><S17D2><S17A2><S17B7>..<S17B8>";
"<BASE><VRNT1><BASE><BASE><VRNT1><BASE>";"<MIN><MIN><MIN><MIN><MIN><MIN>";
<SW_17CC_17A5>..<SW_17CC_17A6> % KHMER INDEPENDENT VOWEL QI, KHMER SIGN
ROBAT..KHMER INDEPENDENT VOWEL QII, KHMER SIGN ROBAT

<SW_17CC_17A7>
"<S179A><S17D2><S17A2><S17BB>";"<BASE><VRNT1><BASE><BASE><VRNT1><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17A7>

% KHMER INDEPENDENT VOWEL QU, KHMER SIGN ROBAT

<SW_17CC_17A8>
"<S179A><S17D2><S17A2><S17BB>";"<BASE><VRNT1><BASE><BASE><VRNT2><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17A8>

% KHMER INDEPENDENT VOWEL QUK;  KHMER SIGN ROBAT

<SW_17CC_17A9>
"<S179A><S17D2><S17A2><S17BC>";"<BASE><VRNT1><BASE><BASE><VRNT1><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17A9>

% KHMER INDEPENDENT VOWEL QUU;  KHMER SIGN ROBAT

<SW_17CC_17AA>
"<S179A><S17D2><S17A2><S17BC>";"<BASE><VRNT1><BASE><BASE><VRNT2><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17AA>

% KHMER INDEPENDENT VOWEL QUUV;  KHMER SIGN ROBAT

<SW_17CC_17AF>..<SW_17CC_17B1>           "<S179A><S17D2><S17A2><S17C2>..<S17C4>";
"<BASE><VRNT1><BASE><BASE><VRNT1><BASE>";"<MIN><MIN><MIN><MIN><MIN><MIN>";
<SW_17CC_17AF>..<SW_17CC_17B1> % KHMER INDEPENDENT VOWEL QE, KHMER SIGN
ROBAT..KHMER INDEPENDENT VOWEL QOO TYPE ONE, KHMER SIGN ROBAT

<SW_17CC_17B2>
"<S179A><S17D2><S17A2><S17C4>";"<BASE><VRNT1><BASE><BASE><VRNT2><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17B2>

% KHMER INDEPENDENT VOWEL QOO TYPE TWO;  KHMER SIGN ROBAT
```

Deleted: /Thai/Lao

```
<SW_17CC_17B3>
"<S179A><S17D2><S17A2><S17C5>";"<BASE><VRNT1><BASE><BASE><VRNT1><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17B3>

% KHMER INDEPENDENT VOWEL QAU;  KHMER SIGN ROBAT

%%% Khmer OM and AAM (the NIKAHIT should be written after the vowel):

<U17BB_17C6> <S17BB_17C6>;<BASE>;<MIN>;<U17BB_17C6> % KHMER VOWEL SIGN U, KHMER
SIGN NIKAHIT

<U17C6_17BB> <S17BB_17C6>;<BASE>;<MIN>;<U17C6_17BB> % KHMER SIGN NIKAHIT, KHMER
VOWEL SIGN U

<U17B6_17C6> <S17B6_17C6>;<BASE>;<MIN>;<U17B6_17C6> % KHMER VOWEL SIGN AA, KHMER
SIGN NIKAHIT

<U17C6_17B6> <S17B6_17C6>;<BASE>;<MIN>;<U17C6_17B6> % KHMER SIGN NIKAHIT, KHMER
VOWEL SIGN AA

reorder-end
```

```
Deleted: %%% Thai pre-
vowel/consonant swaps (note that
SARA AE is ordered as <SARA E, SARA
E>):¶
<U0E40_0E01>..<U0E40_0E2E>
"<S0E01>..<S0E2E><S0E40>";"<BASE><B
ASE>";"<MIN><MIN>";
<U0E40_0E01>..<U0E40_0E2E> % E, KO
KAI..E, HO NOHUK¶
<U0E40_0E40_0E01>..<U0E40_0E40_0E2E
>
"<S0E01>..<S0E2E><S0E40><S0E40>";"<
BASE><BASE><BASE>";
"<MIN><MIN><MIN>";<U0E40_0E40_0E01>
..<U0E40_0E40_0E2E> % E, E, KO KAI..
E, E, HO NOHUK¶
<U0E41_0E01>..<U0E41_0E2E>
"<S0E01>..<S0E2E><S0E40><S0E40>";"<
BASE><BASE><BASE>";
"<MIN><MIN><MIN>";<U0E41_0E01>..<U0
E41_0E2E> % AE, KO KAI..AE, HO
NOHUK¶
<U0E42_0E01>..<U0E42_0E2E>
"<S0E01>..<S0E2E><S0E42>";"<BASE><B
ASE>";"<MIN><MIN>";
<U0E42_0E01>..<U0E42_0E2E> % O, KO
KAI..O, HO NOHUK¶
<U0E43_0E01>..<U0E43_0E2E>
"<S0E01>..<S0E2E><S0E43>";"<BASE><B
ASE>";"<MIN><MIN>";
<U0E43_0E01>..<U0E43_0E2E> % AI
MAIMUAN, KO KAI..AI MAIMUAN, HO
NOHUK¶
<U0E44_0E01>..<U0E44_0E2E>
"<S0E01>..<S0E2E><S0E44>";"<BASE><B
ASE>";"<MIN><MIN>";
<U0E44_0E01>..<U0E44_0E2E> % AI
MAIMALAI, KO KAI..AI MAIMALAI, HO
NOHUK¶
%%% Lao pre-vowel/consonant swap
(note that VOWEL EI is ordered as
<VOWEL E, VOWEL E>):¶
%%% (The code point ranges contain
some code points which are not
assigned to any character.¶
%%% However, the (level 1) weights
should be declared and
appropriately weighted.)¶
<U0EC0_0E81>..<U0EC0_0EAE>
"<S0E81>..<S0EAE><S0EC0>";"<BASE><B
ASE>";"<MIN><MIN>";
<U0EC0_0E81>..<U0EC0_0EAE> % E,
KO..E, HO TAM¶                ... [2]
```

```
Deleted: %%% Thai SARA AM is ordered
as <SARA AA, NIKHAHIT>, but can be
written the other way around:¶
<U0E4D_0E32>
"<S0E32><S0E4D>";"<BASE><BASE>";"<M
IN><MIN>";<U0E4D_0E32> ¶
% THAI CHARACTER NIKHAHIT, THAI
CHARACTER SARA AA¶
```

```
Deleted: %%% Lao VOWEL AM is ordered
as <VOWEL AA, NIGGAHITA>, but can
be written the other way around:¶
<U0EB3>
"<S0EB3>";"<BASE>";"<COMPAT>";<U0EB
3> % LAO VOWEL SIGN AM ¶
<U0EB2_0ECD>
"<S0EB3>";"<BASE>";"<MIN>";<U0EB2_0
ECD> % LAO VOWEL SIGN AA, LAO
NIGGAHITA¶
<U0ECD_0EB2>
"<S0EB3>";"<BASE>";"<MIN>";<U0ECD_0
EB2> % LAO NIGGAHITA, LAO VOWEL
SIGN AA¶
```

Page 25, Annex C

Replace section C.2 with the following:

C.2. Thai string ordering

This annex explains some of the principles behind the tailoring of the CTT given in annex B.5 above, as well as the CTT ordering for Thai (and to some extent Lao).

C.2.1. Thai ordering principles

The widely accepted standard for Thai lexicographical ordering is defined in the Royal Institute Dictionary 2542 B.E. Edition (1999 A.D.), the official standard Thai dictionary. The ordering principles are:

Deleted: 2525 B.E. Edition (1982 A.D.)

- Words are ordered alphabetically, not phonetically. Consonants order is:

  ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท

  ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ฤ ล ฦ ว ศ ษ ส ห ฬ อ ฮ

- Vowels, and nikhahit, are also ordered by written forms, not by sounds. Vowels and nikhahit order is:

  -ํ -ะ -ั -า -ๅ -ำ -ิ -ี -ึ -ื -ุ -ู เ- แ - โ - ใ - ไ - -ฺ

  อ ว ย are always ordered as consonants, although they sometimes act as vowels.

  ๅ is a long-legged variant of า , used with the long-legged consonants ฤ and ฦ : ฤ ๅ and ฦ ๅ .

  ำ is logically an า followed by a ํ. However, the Unicode compatibility decomposition of the precomposed character is to a followed by an า , so this misspelling must be handled as well.

- No syllable structure or word boundary analysis is required, as Thai lexicons are ordered alphabetically, not phonetically. Note that Thai normally does not use any word separator, except, and exceptionally, zero width space.

- String comparison is performed from left to right, but considering initial consonants before vowels in the same syllable. Leading vowels (เ- แ- โ- ใ- ไ-, corresponding to characters U+0E40-U+0E44), which are written before the consonant, must be considered after the consonant. Therefore, rearrangement (in some way) is needed before comparison.

- Tones and diacritics are ignored at level 1. At level 2 their order is:

  -็ -์ -๎ -่ -้ -๊ -๋

- Since Thai, Lao, and Khmer are uncased, it may seem that the third level is not needed for Thai, Lao, and Khmer string ordering. However, the third level is used to differentiate LAKKHANGYAO (ๅ) as a variant of SARA AA (า), since similar variants in other scripts are differentiated in a like manner at level 3, as well as for other variation cases in Thai, Lao, and Khmer.

- When Thai punctuation marks (๛ ๚ ฯ ๏ ๆ) are concerned, another level of weights is required for them. This corresponds to the fourth level in the Common Template Table. In string ordering, punctuation marks are less significant than any tone marks and diacritics, and must be ignored in all the first three levels. Note that PAIYANNOI (ฯ) and THAI CHARACTER MAIYAMOK (repeat mark ๆ) are regarded and ordered as punctuation marks, not letters, despite their Unicode general category as “Lo” and “Lm” respectively. For example, “ข้างๆ, ข้างกบ, ข้างๆ คูๆ, ข้างจัน” is a valid order in the Royal Institute Dictionary. In the first level, the considered weights correspond to ขาง, ขางกบ, ขางคู, and ขางจัน respectively.

Deleted: ๏

Deleted: repeat mark

- The ten Thai decimal digits (๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙), each semantically equivalent to Arabic digit 0-9, respectively. Their weights are then equal to their corresponding Arabic digit in the first level, and are different in the second level, to distinguish script.

C.2.2. Vowel/consonant rearrangement

Regarding the handling of pre-vowels, either a collation preparation or collating-element grouping (as in the tailoring in annex B.5 above) is required. The collation preparation scans the string once and swaps every leading Thai vowel with its succeeding character (ideally only if the succeeding character is a Thai consonant). The prepared string is then passed to the normal weight calculation process. Another way to manage this is by means of collating-element formation – the approach taken both by the CTT of this standard and by the collation weighting table of the Unicode Collation Algorithm (UTS #10). Every possible pair of leading vowel and consonant is defined as a collating-element, whose weight equals to that of the rearranged substring. In addition, since two เ in sequence look just like a แ , two เ in sequence should be handled just like a แ .

Deleted: (as in UAX 10)

Deleted: s

Note that the rearrangement of each leading vowel is simply performed with its immediate succeeding consonant. No consonant cluster analysis is needed. Indeed, doing so would result in ambiguities or yield a different order than that specified in the Royal Institute Dictionary. For example:

1. Ambiguities: The problem with ambiguity is illustrated by the word “เพลา”. It has two potential pronunciations: either as a two-syllable word, “phe-la” (meaning “time”), or as a one-syllable word, “phlao” (meaning “axle” or “abate”). A rearrangement algorithm which follows the distinct pronunciation of the potential cluster ‘พล’ in this string would result in distinct keys, “พเลา” and “พลเา”, and therefore different weights, which are equally legal. Both words need to have the same weight to be sortable, however.

2. Non-conforming ordering: To illustrate the difference in ordering caused by the treatment of consonant clusters, consider these words, shown in conforming order: “เพล, เพลง, เพศ”. The correct rearrangement ignores any clusters and results in the following: “พเล, พเลง, พเศ”, which sorts in the order shown. If, however, pairs of consonants that form legal clusters were grouped as single collation elements (regardless of actual pronunciation where the potential pronunciation is ambiguous), then the results of rearrangement would be “<พล>เ, <พล>เง, พเศ”, which would yield the (non-conforming) ordering “เพศ, เพล, เพลง”. Again, if actual clusters were grouped as single collation elements (with some disambiguation effort), then the results of rearrangement would be “พเล, <พล>เง, พเศ”, which would yield the (non-conforming) ordering “เพล, เพศ, เพลง”.

C.2.3. Example ordered strings

Here is an ordering example: Example for Thai (sorted order)			
ก ก	โ ก น	แข่งขัน	ผัด
ก ร ร ม	โกร๋น	แ ข น	ฯ พ ณฯ
กรรม์	ใกล้	ค ร ร ภ -	พณิชย์
-กระแย่ง	ไก่	ครรภ์	ย่อง
ก ร า บ	ไ ก ล	จุมพล	ร อ ง
กะเกณฑ์	ข้น	จุพล	ฤทธิ์
กัก	ข น า บ	ช า ย	ฤษี
ก้าว	ข้าง	เฒ่า	ฤๅษี
กํ า	ข้างๆ	เ ณร	ลลิตา
กิน	ข้างกระดาน	ต ล า ด	ฦ ๅ ช า
กี่	ข้างขึ้น	ทูลเกล้า	ว ก
กึ๋น	ข้างควาย	ทูลเกล้าฯ	ศ า ล
กุน	ข้างๆ คูๆ	ทูลเกล้าทูลกระหม่อม	หริภุญชัย
กูด	ข้างเงิน	น้า	หฤทัย
เก้ง	ข้างออก	น้ำ	ห ล ง
เกล้า	เ ข น	นี้	แหง่
เกลียว	เข็น	บุญหลง	แห่ง
เก้า	เข่น	บุญ-ห ล ง	แ ห น ม
เ ก า ะ	เข็ด	ป า	แ ห น ห ว ง
เกี่ยว	แข็ง	ป่า	แ ห บ
เกี๊ยะ	แข่ง	ป้า	แ ห ม
เกือก	แข้ง	ป๊า	อ า น
แ ก ง	แข้งขวา	ป๋า	ฮ า
แ ก ะ	แข็งขัน	ป า น	

Bibliography

Replace bibliography by the following :

Bibliography

The following standards and documents are considered relevant to this standard, in addition to the normative references.

- *ISO/IEC 10646-1:1993/Amd.9:1997 Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane Amendment 9: Identifiers for characters.*
- CAN/CSA Z243.4.1-1998 – *Canadian Alphanumeric Ordering Standard, A National Standard of Canada, Canadian Standards Association*.
- CAN/CSA Z243.230-1998 – *Minimum Canadian Software Localisation Conventions, A National Standard of Canada*.
- DS 377:1980 – *Alfabetiseringsregler*, Dansk Standard.
- Gavare, Rolf, *Alphabetic ordering in a lexicological perspective, Studies in Computer-Aided Lexicology*, 1988, pp. 63–102.
- ISO/IEC 646, *Information technology – ISO 7-bit coded character set for information interchange*.
- ISO/IEC 2022, *Information technology – Character code structure and extension techniques*.
- ISO/IEC 6937, *Information technology – Coded graphic character set for text communication – Latin alphabet*.
- ISO/IEC 8859-1, *Information technology – 8-bit single-byte coded graphic character sets – Part 1: Latin alphabet No. 1*.
- ISO/IEC 8859-2, *Information technology – 8-bit single-byte coded graphic character sets – Part 2: Latin alphabet No. 2*.
- ISO/IEC 8859-3, *Information technology – 8-bit single-byte coded graphic character sets – Part 3: Latin alphabet No. 3*.
- ISO/IEC 8859-4, *Information technology – 8-bit single-byte coded graphic character sets – Part 4: Latin alphabet No. 4*.
- ISO/IEC 8859-5, *Information technology – 8-bit single-byte coded graphic character sets – Part 5: Latin/Cyrillic alphabet*
- ISO/IEC 8859-6, *Information technology – 8-bit single-byte coded graphic character sets – Part 6: Latin/Arabic alphabet*.
- ISO/IEC 8859-7, *Information technology – 8-bit single-byte coded graphic character sets – Part 7: Latin/Greek alphabet*.
- ISO/IEC 8859-8, *Information technology – 8-bit single-byte coded graphic character sets – Part 8: Latin/Hebrew alphabet*.
- ISO/IEC 8859-9, *Information technology – 8-bit single-byte coded graphic character sets – Part 9: Latin alphabet No. 5*.
- ISO/IEC 8859-10, *Information technology – 8-bit single-byte coded graphic character sets – Part 10: Latin alphabet No. 6*.

- ISO/IEC 8859-13, *Information technology – 8-bit single-byte coded graphic character sets – Part 13: Part 13: Latin alphabet No. 7*.

- ISO/IEC 8859-14, *Information technology – 8-bit single-byte coded graphic character sets – Part 14: Latin alphabet No. 8 (Celtic)*.

- ISO/IEC 8859-15, *Information technology – 8-bit single-byte coded graphic character sets – Part 15: Latin alphabet No. 9*.

- ISO/IEC 9945-2, *Information technology – Portable Operating System Interface (POSIX) - Part 2: Shell and Utilities*.

- ISO/IEC DTR 14652, *Information technology – Specification method for cultural conventions*.

- *Règles du classement alphabétique en langue française et procédure informatisée pour le tri, Conseil du trésor du Québec*, URL: http://www.tresor.gouv.qc.ca/doc/classm.htm.

- *Retskrivningsordbogen – 2nd edition 1996*, Dansk Sprognævn & Aschehoug Dansk Forlag A/S.

- *Technique de réduction - Tris informatiques à quatre clés, Conseil du trésor du Québec*, URL: http://www.tresor.gouv.qc.ca/doc/techtri.htm.

- *Teknisk norm nr. 34, Swedish Alphanumeric Sorting, Statskontoret, 1992*. (Includes Gavare's paper as an annex.)

Formatted

- *The Unicode Standard, Version 4.0, The Unicode Consortium, Addison-Wesley, 2003. ISBN 0-321-18578-1, as amended by Unicode 4.0.1 (http://www.unicode.org/versions/Unicode4.0.1/) and Unicode 4.1.0 (http://www.unicode.org/versions/Unicode4.1.0).*

Formatted: Bullets and Numbering

- *Unicode Technical Report no. 10, Unicode Collation Algorithm*, The Unicode Consortium, URL: http://www.unicode.org/unicode/reports/tr10/.

Deleted: *The Unicode Standard, Version 2.0, The Unicode Consortium, Addison Wesley Developers Press, ISBN 0-201-48345-9.*¶

Deleted: <#>*Unicode Technical Report no. 8, The Unicode Standard, Version 2.1*, The Unicode Consortium, URL: http://www.unicode.org/unicode/reports/tr8/.¶

- *Unicode Technical Report no. 15, Unicode Normalization Forms*, The Unicode Consortium, URL: http://www.unicode.org/unicode/reports/tr15/.

Page 3: [1] Deleted **Alain LaBonté** **9/2/2005 2:38:00 PM**

```
% Thai:

collating-element <U0E5A_0E30> from "<U0E5A><U0E30>" % special end
symbol

   % Thai pre-vowels:

collating-element <U0E40_0E01>..<U0E40_0E2E> from
"<U0E40><U0E01>..<U0E2E>"

collating-element <U0E40_0E40_0E01>..<U0E40_0E40_0E2E> from
"<U0E40><U0E40><U0E01>..<U0E2E>"

collating-element <U0E41_0E01>..<U0E41_0E2E> from
"<U0E41><U0E01>..<U0E2E>"

collating-element <U0E42_0E01>..<U0E42_0E2E> from
"<U0E42><U0E01>..<U0E2E>"

collating-element <U0E43_0E01>..<U0E43_0E2E> from
"<U0E43><U0E01>..<U0E2E>"

collating-element <U0E44_0E01>..<U0E44_0E2E> from
"<U0E44><U0E01>..<U0E2E>"



collating-element <U0E4D_0E32> from "<S0E4D><S0E32>" % AM (in the wrong
order really)



% Lao:

   % Lao pre-vowels:

collating-element <U0EC0_0E81>..<U0EC0_0EAE> from
"<U0EC0><U0E81>..<U0EAE>"

collating-element <U0EC0_0EC0_0E81>..<U0EC0_0EC0_0EAE> from
"<U0EC0><U0EC0><U0E81>..<U0EAE>"

collating-element <U0EC1_0E81>..<U0EC1_0EAE> from
"<U0EC1><U0E81>..<U0EAE>"

collating-element <U0EC2_0E81>..<U0EC2_0EAE> from
"<U0EC2><U0E81>..<U0EAE>"

collating-element <U0EC3_0E81>..<U0EC3_0EAE> from
"<U0EC3><U0E81>..<U0EAE>"

collating-element <U0EC4_0E81>..<U0EC4_0EAE> from
"<U0EC4><U0E81>..<U0EAE>"
```

```
   % LAO HO NO

collating-element <U0EC0_0EDC> from "<U0EC0><U0EDC>"

collating-element <U0EC0_0EC0_0EDC> from "<U0EC0><U0EC0><U0EDC>"

collating-element <U0EC1_0EDC> from "<U0EC1><U0EDC>"

collating-element <U0EC2_0EDC> from "<U0EC2><U0EDC>"

collating-element <U0EC3_0EDC> from "<U0EC3><U0EDC>"

collating-element <U0EC4_0EDC> from "<U0EC4><U0EDC>"

   % LAO HO MO

collating-element <U0EC0_0EDD> from "<U0EC0><U0EDD>"

collating-element <U0EC0_0EC0_0EDD> from "<U0EC0><U0EC0><U0EDD>"

collating-element <U0EC1_0EDD> from "<U0EC1><U0EDD>"

collating-element <U0EC2_0EDD> from "<U0EC2><U0EDD>"

collating-element <U0EC3_0EDD> from "<U0EC3><U0EDD>"

collating-element <U0EC4_0EDD> from "<U0EC4><U0EDD>"

   % U+0EAB LAO LETTER HO SUNG, U+200D ZERO WIDTH JOINER, U+0EA5 LAO
LETTER LO LOOT

collating-element              <U0EC0_0EAB_200D_0EA5>              from
"<U0EC0><U0EAB><U200D><U0EA5>"

collating-element           <U0EC0_0EC0_0EAB_200D_0EA5>            from
"<U0EC0><U0EC0><U0EAB><U200D><U0EA5>"

collating-element              <U0EC1_0EAB_200D_0EA5>              from
"<U0EC1><U0EAB><U200D><U0EA5>"

collating-element              <U0EC2_0EAB_200D_0EA5>              from
"<U0EC2><U0EAB><U200D><U0EA5>"

collating-element              <U0EC3_0EAB_200D_0EA5>              from
"<U0EC3><U0EAB><U200D><U0EA5>"

collating-element              <U0EC4_0EAB_200D_0EA5>              from
"<U0EC4><U0EAB><U200D><U0EA5>"

collating-element <U0ECD_0EB2> from "<S0ECD><S0EB2>" % AM (in the wrong
order really)

collating-element  <U0EB2_0ECD>  from  "<S0EB2><S0ECD>"  %  AM  (properly
spelled)
```

Page 6: [2] Deleted **Alain LaBonté** **9/2/2005 2:52:00 PM**

```
%%% Thai pre-vowel/consonant swaps (note that SARA AE is ordered as
<SARA E, SARA E>):

<U0E40_0E01>..<U0E40_0E2E>
"<S0E01>..<S0E2E><S0E40>";"<BASE><BASE>";"<MIN><MIN>";
<U0E40_0E01>..<U0E40_0E2E> % E, KO KAI..E, HO NOHUK

<U0E40_0E40_0E01>..<U0E40_0E40_0E2E>
"<S0E01>..<S0E2E><S0E40><S0E40>";"<BASE><BASE><BASE>";
"<MIN><MIN><MIN>";<U0E40_0E40_0E01>..<U0E40_0E40_0E2E> % E, E, KO KAI..
E, E, HO NOHUK

<U0E41_0E01>..<U0E41_0E2E>
"<S0E01>..<S0E2E><S0E40><S0E40>";"<BASE><BASE><BASE>";
"<MIN><MIN><MIN>";<U0E41_0E01>..<U0E41_0E2E> % AE, KO KAI..AE, HO NOHUK

<U0E42_0E01>..<U0E42_0E2E>
"<S0E01>..<S0E2E><S0E42>";"<BASE><BASE>";"<MIN><MIN>";
<U0E42_0E01>..<U0E42_0E2E> % O, KO KAI..O, HO NOHUK

<U0E43_0E01>..<U0E43_0E2E>
"<S0E01>..<S0E2E><S0E43>";"<BASE><BASE>";"<MIN><MIN>";
<U0E43_0E01>..<U0E43_0E2E> % AI MAIMUAN, KO KAI..AI MAIMUAN, HO NOHUK

<U0E44_0E01>..<U0E44_0E2E>
"<S0E01>..<S0E2E><S0E44>";"<BASE><BASE>";"<MIN><MIN>";
<U0E44_0E01>..<U0E44_0E2E> % AI MAIMALAI, KO KAI..AI MAIMALAI, HO NOHUK

%%% Lao pre-vowel/consonant swap (note that VOWEL EI is ordered as
<VOWEL E, VOWEL E>):

%%% (The code point ranges contain some code points which are not
assigned to any character.

%%% However, the (level 1) weights should be declared and appropriately
weighted.)

<U0EC0_0E81>..<U0EC0_0EAE>
"<S0E81>..<S0EAE><S0EC0>";"<BASE><BASE>";"<MIN><MIN>";
<U0EC0_0E81>..<U0EC0_0EAE> % E, KO..E, HO TAM

<U0EC0_0EC0_0E81>..<U0EC0_0EC0_0EAE>
"<S0E81>..<S0EAE><S0EC0><S0EC0>";"<BASE><BASE><BASE>";
"<MIN><MIN><MIN>";<U0EC0_0EC0_0E81>..<U0EC0_0EC0_0EAE> % E, E, KO..E, E,
HO TAM

<U0EC1_0E81>..<U0EC1_0EAE>
"<S0E81>..<S0EAE><S0EC0><S0EC0>";"<BASE><BASE><BASE>";
"<MIN><MIN><MIN>";<U0EC1_0E81>..<U0EC1_0EAE> % EI, KO, EI, HO TAM
```

```
<U0EC2_0E81>..<U0EC2_0EAE>
"<S0E81>..<S0EAE><S0EC2>";"<BASE><BASE>";"<MIN><MIN>";
<U0EC2_0E81>..<U0EC2_0EAE> % O, KO..O, HO TAM

<U0EC3_0E81>..<U0EC3_0EAE>
"<S0E81>..<S0EAE><S0EC3>";"<BASE><BASE>";"<MIN><MIN>";
<U0EC3_0E81>..<U0EC3_0EAE> % AY, KO..AY, HO TAM

<U0EC4_0E81>..<U0EC4_0EAE>
"<S0E81>..<S0EAE><S0EC4>";"<BASE><BASE>";"<MIN><MIN>";
<U0EC4_0E81>..<U0EC4_0EAE> % AI, KO..AI, HO TAM

<U0EC0_0EDC>
"<S0EAB><S0E99><S0EC0>";"<BASE><BASE><BASE>";"<COMPAT><COMPAT><COMPAT>";

<U0EC0_0EDC> % E, HO NO

<U0EC0_0EC0_0EDC>
"<S0EAB><S0E99><S0EC0><S0EC0>";"<BASE><BASE><BASE><BASE>";
"<COMPAT><COMPAT><COMPAT><COMPAT>";<U0EC0_0EC0_0EDC> % E, E, HO NO

<U0EC1_0EDC> "<S0EAB><S0E99><S0EC0><S0EC0>";"<BASE><BASE><BASE><BASE>";

"<COMPAT><COMPAT><COMPAT><COMPAT>"; <U0EC1_0EDC> % EI, HO NO

<U0EC2_0EDC>
"<S0EAB><S0E99><S0EC2>";"<BASE><BASE><BASE>";"<COMPAT><COMPAT><COMPAT>";

<U0EC2_0EDC> % O, HO NO

<U0EC3_0EDC>
"<S0EAB><S0E99><S0EC3>";"<BASE><BASE><BASE>";"<COMPAT><COMPAT><COMPAT>";

<U0EC3_0EDC> % AY, HO NO

<U0EC4_0EDC>
"<S0EAB><S0E99><S0EC4>";"<BASE><BASE><BASE>";"<COMPAT><COMPAT><COMPAT>";

<U0EC3_0EDC> % AI, HO NO

<U0EC0_0EDD>
"<S0EAB><S0EA1><S0EC0>";"<BASE><BASE><BASE>";"<COMPAT><COMPAT><COMPAT>";

<U0EC0_0EDD> % E, HO MO

<U0EC0_0EC0_0EDD>
"<S0EAB><S0EA1><S0EC0><S0EC0>";"<BASE><BASE><BASE><BASE>";
"<COMPAT><COMPAT><MIN><MIN>";<U0EC0_0EC0_0EDD> % E, E, HO MO

<U0EC1_0EDD>  "<S0EAB><S0EA1><S0EC0><S0EC0>";"<BASE><BASE><BASE><BASE>";
"<COMPAT><COMPAT><COMPAT><COMPAT>";<U0EC1_0EDD> % EI, HO MO

<U0EC2_0EDD>
"<S0EAB><S0EA1><S0EC2>";"<BASE><BASE><BASE>";"<COMPAT><COMPAT><COMPAT>";
```

```
<U0EC2_0EDD> % O, HO MO

<U0EC3_0EDD>
"<S0EAB><S0EA1><S0EC3>";"<BASE><BASE><BASE>";"<COMPAT><COMPAT><COMPAT>";

<U0EC3_0EDD> % AY, HO MO

<U0EC4_0EDD>
"<S0EAB><S0EA1><S0EC4>";"<BASE><BASE><BASE>";"<COMPAT><COMPAT><COMPAT>";

<U0EC4_0EDD> % AI, HO MO

%% HO LO is a ligature which does not have its own code point, so ZWJ is
used.

<U0EC0_0EAB_200D_0EA5>
"<S0EAB><S0EA5><S0EC0>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";
<U0EC0_0EAB_200D_0EA5> % E, HO LO,

<U0EC0_0EC0_0EAB_200D_0EA5>
"<S0EAB><S0EA5><S0EC0><S0EC0>";"<BASE><BASE><BASE><BASE>";
"<MIN><MIN><MIN><MIN>";<U0EC0_0EC0_0EAB_200D_0EA5> % E, E, HO LO

<U0EC1_0EAB_200D_0EA5>
"<S0EAB><S0EA5><S0EC0><S0EC0>";"<BASE><BASE><BASE><BASE>";
"<MIN><MIN><COMPAT><COMPAT>";<U0EC1_0EAB_200D_0EA5> % EI, HO LO

<U0EC2_0EAB_200D_0EA5>
"<S0EAB><S0EA5><S0EC2>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";
<U0EC2_0EAB_200D_0EA5> % O, HO LO

<U0EC3_0EAB_200D_0EA5>
"<S0EAB><S0EA5><S0EC3>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";
<U0EC3_0EAB_200D_0EA5> % AY, HO LO

<U0EC4_0EAB_200D_0EA5>
"<S0EAB><S0EA5><S0EC4>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";
<U0EC4_0EAB_200D_0EA5> % AI, HO LO
```