



Proposed Update Unicode Standard Annex #34

UNICODE NAMED CHARACTER SEQUENCES

Version	Unicode 5.1.0 (draft)
Authors	Ken Whistler and Asmus Freytag (ken@unicode.org)
Date	2007-09-24
This Version	http://www.unicode.org/reports/tr34/tr34-6.html
Previous Version	http://www.unicode.org/reports/tr34/tr34-5.html
Latest Version	http://www.unicode.org/reports/tr34/
Revision	6

Summary

This annex defines the concept of Unicode named character sequences, specifies a notational convention for them and a set of rules constraining possible names applied to character sequences.

Status

This is a **draft** document which may be updated, replaced, or superseded by other documents at any time. Publication does not imply endorsement by the Unicode Consortium. This is not a stable document; it is inappropriate to cite this document as other than a work in progress.

A Unicode Standard Annex (UAX) forms an integral part of the Unicode Standard, but is published online as a separate document. The Unicode Standard may require conformance to normative content in a Unicode Standard Annex, if so specified in the Conformance chapter of that version of the Unicode Standard. The version number of a UAX document corresponds to the version of the Unicode Standard of which it forms a part.

Please submit corrigenda and other comments with the online reporting form [[Feedback](#)]. Related information that is useful in understanding this annex is found in Unicode Standard Annex #41, "[Common References for Unicode Standard Annexes](#)." For the latest version of the Unicode Standard, see [[Unicode](#)]. For a list of current Unicode Technical Reports, see [[Reports](#)]. For more information about versions of the Unicode Standard, see [[Versions](#)].

Contents

- 1 [Overview](#)
 - 1.1 [Relation to Variation Sequences](#)
- 2 [Definitions and Notation](#)
- 3 [Conformance](#)
 - 3.1 [Provisional Process for Named Character Sequences](#)
- 4 [Names](#)
- 5 [Data Files](#)
- [Acknowledgments](#)
- [References](#)

[Modifications](#)

1 Overview

The Unicode Standard specifies notational conventions for referring to sequences of characters (or code points), using angle brackets surrounding a comma-delimited list of code points, code points plus character names, and so on. For example, both of the designations in *Table 1* refer to a combining character sequence consisting of the letter “a” with a circumflex and an acute accent applied to it.

Table 1. Example of a Combining Character Sequence

<U+0061, U+0302, U+0301>
<U+0061 LATIN SMALL LETTER A, U+0302 COMBINING CIRCUMFLEX ACCENT, U+0301 COMBINING ACUTE ACCENT>






See *Appendix A, Notational Conventions*, in [\[Unicode\]](#) for the description of the conventions for expression of code points and for the representation of sequences of code points.


The Unicode conventions for referring to a sequence of characters (or code points) are a generalization of the formal syntax specified in ISO/IEC 10646:2003 for UCS Sequence Identifiers, or USI. A USI has the form

<UID₁, UID₂, ... UID_n>

where the UID_i represent the short identifiers for code points—most commonly “U+0061” or “0061”. A USI must contain at least two code points.

Table 2. Examples of Named Sequences

Sequence	Name	Notes on Usage
 012B 0300	LATIN SMALL LETTER I WITH MACRON AND GRAVE	Livonian
 02E5 02E9	MODIFIER LETTER EXTRA-HIGH EXTRA-LOW CONTOUR TONE BAR	Contour tone letter
 31F7 309A	KATAKANA LETTER AINU P	Ainu in kana transcription
 17BB 17C6	KHMER VOWEL SIGN SRAK OM	Khmer
 17B6 17C6	KHMER VOWEL SIGN SRAK AM	Khmer

	KHMER CONSONANT SIGN COENG KA	Khmer
---	-------------------------------	-------

Such a conventional notation for sequences of Unicode code points that are treated as a unit is often useful. For example, other standards may need to refer to entities that are represented in Unicode by sequences of characters. Mapping tables may map single characters in other standards to sequences of Unicode characters, and listings of repertoire coverage for fonts or keyboards may need to reference entities that do not correspond to single Unicode code points.

In some limited circumstances it is necessary to also provide a name for such sequences. The primary example is the need to have an identifier for a sequence to correlate with an identifier in another standard, for which a cross-mapping to Unicode is desired. To address this need, the Unicode Standard defines a mechanism for naming sequences and provides a short list of sequences that have been formally named. This list is deliberately selective: it is neither possible nor desirable to attempt to provide names for all possible sequences of Unicode characters that could be of interest.

This annex defines the concept of a *Unicode named character sequence*, specifies a notational convention for such sequences, and a set of rules constraining possible names applied to character sequences. *Section 5, Data Files*, identifies the data file containing the normative list of Unicode named character sequences. As is the case for character names, named character sequences are strictly synchronized with ISO/IEC 10646.

Table 2 provides some examples of Unicode named character sequences to illustrate the kinds of entities that have been formally named. The “Sequence” column illustrates the entity in question with a representative rendering above the sequence of encoded Unicode characters that represent that entity. The “Name” column shows the name that has been associated with that sequence.

1.1 Relation to Variation Sequences

Unicode named character sequences differ from Unicode variation sequences. The latter are documented in *Section 16.4, Variation Selectors*, in [Unicode] and are listed exhaustively in the data file StandardizedVariants.txt in the Unicode Character Database [UCD].

Variation sequences always consist of a sequence of precisely defined code points, the second of which must be a variation selector. There are additional constraints on which types of characters they can start with. Variation sequences have a restricted range of glyphic shapes, but have no associated name.

Named character sequences can, in principle, consist of code point sequences of any length, without constraints on what types of characters are involved. They do not have a specifically defined glyphic shape, but they *do* have a formally specified name associated with them.

2 Definitions and Notation

SD1 Unicode named character sequence: A specific sequence of two or more Unicode characters, together with a formal name designating that sequence.

The notation for a Unicode named character sequence consists of the general conventions for character sequences in *Appendix A, Notational Conventions*, of [Unicode], together with name conventions as specified in *Section 4, Names*. Thus a typical representation of a Unicode named character sequence would be

<U+012B, U+0300> LATIN SMALL LETTER I WITH MACRON AND GRAVE

In contexts that supply other clear means for delimitation, such as data files or tables, the bracketing and comma delimitation conventions for the sequences may be dropped, as in

012B 0300;LATIN SMALL LETTER I WITH MACRON AND GRAVE

3 Conformance

Conformance to the Unicode Standard *requires* conformance to the specification in this annex. The relationship between conformance to the Unicode Standard and conformance to an individual Unicode Standard Annex (UAX) is described in more detail in *Section 3.2, Conformance Requirements*, in [\[Unicode\]](#).

UAX34-C1: *If a process purports to implement Unicode named character sequences, it shall use only those named character sequences defined in the file NamedSequences.txt in the Unicode Character Database.*

Only the named character sequences in NamedSequences.txt are named in this standard. No other Unicode character sequences are given names in this version of the Unicode Standard, although named character sequences may be added in the future. Only sequences that are in Normalization Form NFC are given names in the Unicode Standard.

Conformance to this clause should not be construed as preventing implementers from providing informal names of their choice to any entities or character sequences, as appropriate. However, such informal names are not specified in any way by this standard for use in interchange.

The use of unnamed sequences is not affected by the specifications in this annex.

3.1 Provisional Process for Named Character Sequences

When named character sequences are first suggested for inclusion in the Unicode Standard, they may be accepted provisionally. In such cases, they are listed in the file NamedSequencesProv.txt. See [\[DataProv\]](#).

Character sequences and proposed names listed in NamedSequencesProv.txt are *provisional* only and have no other status. They become part of the standard itself only when approved for inclusion in NamedSequences.txt.

The use of a provisional list is meant to allow sufficient time for review and comment on proposed named character sequences before they are finally approved. This also enables the normative data file, NamedSequences.txt, to remain stable.

4 Names

Names of Unicode named character sequences are unique. They are part of the same namespace as Unicode character names. As a result, where a name exists as a character name, a modified name must be assigned instead. The same applies to not-yet-encoded characters.

Where possible, the names for sequences are constructed by appending the names of the constituent elements together while eliding duplicate elements, and possibly introducing the words between elements for clarity. Where this process would result in a name that already exists, the name is modified suitably to guarantee uniqueness. *Table 3* gives some examples of names for hypothetical sequences constructed according in this manner.

Table 3. Examples of Hypothetical Sequence Names

USI	Alternate Representation of	Name
-----	-----------------------------	------

	Sequence	
<0041, 0043, 0043>	<A, B, C>	LATIN CAPITAL LETTER A B C
<00CA, 0046>	<AE, F>	LATIN CAPITAL LETTER AE F
<0058, 030A>	<X, COMBINING RING ABOVE>	LATIN CAPITAL LETTER X WITH RING ABOVE

Where names are constructed other than by merging existing character names for the constituent characters of the sequence, convention restricts any additional items to the Latin capital letters A to Z, SPACE, HYPHEN-MINUS, and the digits 0 to 9, provided that a digit is not the first character in a word. This convention makes it possible to turn names into identifiers using straightforward transformations.

Names for named sequences are constructed according to the following rules:

R1: Only Latin capital letters A to Z, digits 0 to 9 (provided that a digit is not the first character in a word), SPACE, and HYPHEN-MINUS are used for writing the names.

R2: Only one name is given to each named sequence, and each named sequence must have a unique name within the namespace that named sequences share with character names.

R3: Like character names, names for sequences are unique if they are different even when SPACE and medial HYPHEN-MINUS characters are ignored, and when the strings "LETTER", "CHARACTER", and "DIGIT" are ignored in comparison of the names.

The following two character names are exceptions to this rule, because they were created before this rule was specified:

116C HANGUL JUNGSEONG OE
1180 HANGUL JUNGSEONG O-E

Examples of unacceptable names that are not unique:

SARATI LETTER AA
SARATI CHARACTER AA

These two names would not be unique if the strings "LETTER" and "CHARACTER" were ignored.

R4: Where possible, names for named sequences are constructed by appending the names of the constituent elements together while eliding duplicate elements, and possibly introducing the words "WITH" or "AND" between elements for clarity. Should this process result in a name that already exists, the name is modified suitably to guarantee uniqueness among character names and names for named sequences.

R5: Where applicable, the rules from Appendix L in ISO/IEC 10646:2003 apply.

Note: Just like character names, the names for sequences may be translated, with the translated names for each language being unique with respect to each other and the corresponding set of translated character names. However, translated names are not restricted to the same limited character set as the English names. Translated names may not be suitable as identifiers without modification.

5 Data Files

A normative data file, NamedSequences.txt, is available consisting of those named sequences defined for this version of [Unicode]. The sequences are listed in the data file in an abbreviated

format. For the location of the data file, see [\[Data34\]](#).

In addition, a provisional data file, NamedSequencesProv.txt, is available containing sequences and names proposed for the standard but not yet approved as part of the normative list of named character sequences. For the location of the data file, see [\[DataProv\]](#).

Acknowledgments

Thanks to [Asmus Freytag](#), Mark Davis and Julie Allen for comments on this annex, including earlier versions.

References

For references for this annex, see Unicode Standard Annex #41, "[Common References for Unicode Standard Annexes](#)."

Modifications

The following summarizes modifications from the previous version of this annex.

Revision 6

- [Updated to Unicode 5.1.0.](#)

Revision 5

- Added text to clarify that named character sequences are in NFC and are synchronized with 10646.
- Added text to rule R4 to indicate that "WITH" and "AND" may be added between elements of a name for a named sequence, for clarity in the resulting name.
- Added section 3.1 and DataProv reference.

Revision 4 being a proposed update, only changes between revisions 3 and 5 are noted here.

Revision 3

- Finished editing for initial publication as part of Unicode 4.1.0.

Revision 2

- Internal draft, revised substantially for initial publication as part of Unicode 4.1.0.

Revision 1

- Initial version

Copyright © 2001-2007 Unicode, Inc. All Rights Reserved. The Unicode Consortium makes no expressed or implied warranty of any kind, and assumes no liability for errors or omissions. No liability is assumed for incidental and consequential damages in connection with or arising out of the use of the information or programs contained or accompanying this technical report. The Unicode [Terms of Use](#) apply.

Unicode and the Unicode logo are trademarks of Unicode, Inc., and are registered in some jurisdictions.

