L2/20-009

# Proposal to adapt TUS to extended support for abbreviations

For consideration by Unicode Technical Committee

2020-01-06
Marcel Schneider (charupdate@orange.fr)

*We should always say what we see.*
*Above all we should always*
*—which is most difficult—*
*see what we see.*

Charles Péguy

This proposal adds to the response to Action item 161-A1 as it is complementary to *Proposal to extend support for abbreviations,* submitted simultaneously.

By coincidence, this proposal is also part of Unicode 13.0 beta feedback.

**Note about colors:**

Highlighting is yellow for new text, lime green for reused, and ~~purple & barred~~ for deleted. That color scheme aims at distinguishing moved, copy-pasted or case-converted strings, from those that are added from scratch. Using another color for deletions (plus line through) is for easier fast-reading.

## 1 Section 3.3, definition D5, p. 88

**Change from:**

Character name aliases are also assigned to provide string identifiers for control codes and to recognize widely used alternative names and abbreviations for control codes, format characters and other special-use characters.

**Change to:**

Character name aliases are also assigned to provide string identifiers for control codes and to recognize ~~widely used~~ alternative names and abbreviations for control codes, format characters and other special-use characters.

**Rationale:**

The Core Specification assumes that all standard abbreviations are “widely used” or “commonly occurring”. That is problematic both with respect to less-than-frequent standard abbreviations, and when it comes to standardizing more abbreviations for convenience.

A number of standard abbreviations are not widely used, for example “BS” for “backspace”, since the one widely used abbreviation is “BKSP”, and another less frequent one is “BSP”. “BS” can even be considered ill-formed, as it is lacking the P of the initial “sp”. Consistently, all other Unicode abbreviations involving the word “space” contain “SP”. Some other abbreviations are in NameAliases.txt because they are in ISO 6429, which does not forcibly make them “widely used.”

On the other side, by making this assumption, TUS precludes many abbreviations from becoming standard, such as “OGSP” for OGHAM SPACE MARK, an abbreviation that was supported by the US National Body and UTC in 2008. Most other abbreviations already in—or extrapolated from—dashed- box glyphs in the Code Charts are non-standard even when consisting only of Latin capitals and thus complying to the Unicode character namespace constraints. As a result, they cannot be used anywhere without raising some eyebrows, compelling UTC to rule out such useful an abbreviation as “NBHY”, and expanding string descriptors like “<SHY, NBHY>” to “<SHY, NON-BREAKING HYPHEN>”.

As a consequence I’d suggest leaving a door open for missing abbreviations to enter the Standard

## 2 Section 4.8, Table 4-7., p. 181

**Change from:**

Commonly occurring abbreviations (or acronyms) for control codes, format characters, spaces, and variation selectors

**Change to:**

~~Commonly occurring a~~Abbreviations (~~or~~ may be acronyms or initialisms) for ~~control codes, format characters,~~ various blank or special-use characters such as spaces~~, and~~ or variation selectors

**Rationale:**

The word “acronym” is neither an alternative nor a synonym of “abbreviation”. It is not an alternative since an acronym is a form of abbreviation. It is not a synonym neither because it designates a subset, namely abbreviations that are currently pronounced as if they were a word, as opposed to initialisms, that are pronounced one letter at a time (per TUS, p. 799). Hence both should be mentioned, if any, and “may be” added if “abbreviation” is not the sum of “acronym” and “initialism”; some may be neither of both of these.

The enumeration “control codes, format characters, spaces, and variation selectors” is already incomplete as it doesn’t catch the COMBINING GRAPHEME JOINER, that is neither a control, nor a format character, nor a space nor a variation selector, but a nonspacing combining mark; yet CGJ is in NameAliases.txt. It will grow more incomplete when a set of abbreviations currently missing from NameAliases.txt will be added. Hence this enumeration would better be replaced with an open-ended outline while leaving some examples, in synch with the changes suggested for NameAliases.txt in *Proposal to extend support for abbreviations.*

## 3 Section 4.8, p. 182

**Change from:**

Additional character name aliases match existing and widely used abbreviations (or acronyms) for control codes and for Unicode format characters:

**Change to:**

Additional character name aliases match ~~existing and widely used~~ abbreviations (~~or~~ may be acronyms or initialisms) for various blank or special-use characters such as control codes, ~~and for Unicode~~ format characters or spaces:

**Rationale:**

This is covered by the preceding rationales, except that the enumeration here is even far less complete as it doesn't mention anything more than control characters and format characters, not mentioning spaces any longer, nor variation selectors.

## 4 Section 24.1, p. 909

**Change from:**

Normative aliases which represent commonly used abbreviations for control codes or format characters are shown in all caps, enclosed in parentheses. In contrast, informative aliases are shown in lowercase. For the definitive list of normative aliases, also including their type and suitable for machine parsing, see NameAliases.txt in the UCD.

**Change to:**

Normative aliases which represent commonly used abbreviations for control codes or format characters are shown in all caps.~~,~~ When following an informative alias, they are enclosed in parentheses. In contrast, informative aliases are shown in lowercase. For the ~~definitive~~ standardized list of normative aliases, also including their type and suitable for machine parsing, see NameAliases.txt in the UCD.

**Rationale:**

Parenthesizing abbreviations, even standard ones, is not mandatory in the Code Charts. The statement made does only apply to control codes, not to format characters. Many of these have an annotation containing the word “abbreviated” and no parentheses around the abbreviation, for example SOFT HYPHEN: “commonly abbreviated as SHY”.

The word “definitive” is not backed by any stability policy freezing NameAliases.txt in its actual state. For instance, version Unicode 6.0.0 (2010) was restricted to then-formal aliases (11 entries: LATIN CAPITAL LETTER GHA, LATIN SMALL LETTER GHA, KANNADA LETTER LLLA, LAO LETTER FO FON, LAO LETTER FO FAY, LAO LETTER RO, LAO LETTER LO, TIBETAN MARK BKA- SHOG GI MGO RGYAN, YI SYLLABLE ITERATION MARK, PRESENTATION FORM FOR VERTICAL RIGHT WHITE LENTICULAR BRACKET and BYZANTINE MUSICAL SYMBOL FTHORA SKLIRON CHROMA VASIS). The actual state of the art dates from 2014 (Unicode 7.0.0). Even though since then, the only additions were 11 corrections of character names, the list of normative aliases is by no means “definitive”.

## Acknowledgments

Thanks to everyone who directly or indirectly helped put this paper together.

Thanks to Microsoft for Word Online, OneDrive, VS Code and MSKLC.

Thanks to Google for Google Chrome, Google Search, Google Books, Google Translate and Gmail.