L2/17-324

Two distinct code points:
DECIMAL SEPARATOR and FULL STOP

Dario Schiavon, 2017-09-08

Introduction

Unicode, being an extension of ASCII, inherited a great historical mistake, namely the use of the same code point for characters that sometimes are visually identical (or near-identical) but have completely different meanings.

One example is HYPHEN-MINUS (U+002D), '-', which since the very beginning has been used as both a word separator (i.e., a hyphen, as in "bad-tempered") and a mathematical symbol (i.e., the minus sign, as in "2 - 1"). This problem was recognized by the Unicode committee, which decided to keep the ambiguous ASCII code point, but also added two new code points with the more specific meanings: HYPHEN (U+2010), '-', and MINUS SIGN (U+2212), '−'.

Now, this solution does not go all the way through, because the minus sign also has multiple meanings in the mathematical domain itself (it can be used as a prefix operator, "−1", denoting opposite numbers, or as a binary infix operator, "2 − 1", denoting subtraction). But at least the ambiguity stays within mathematics and does not cross multiple incorrelated domains such as mathematics and English orthography.

Another character with a double meaning is FULL STOP (U+002E). This one is used in the western scripts (and surely many other scripts as well) to mark the end of a sentence. But it is also used in mathematics as a decimal point (as in "1.2" being one plus two-tenths).

Proposal

I hereby propose, for consistency reasons but not only (see below), to rename the U+002E code point as "FULL STOP-DECIMAL SEPARATOR" (as an acknowledgement of the double use that it has enjoyed until today) and to add two new code points with the more specific meanings:

- FULL STOP
- DECIMAL SEPARATOR

The code points could be added in the General Punctuation block (in which there still is 1 unassigned code point) or the Supplemental Punctuation block (in which there are 74 unassigned). To promote the use of these code points, they definitely need to be in the basic multilingual plane.

Note that there are no new glyphs introduced by this proposal, yet. However, once the two functions of the period are distinguished, typeface designers may want to produce small variations between the two, so that they better fit in their contexts are so that careful users may distinguish them by eye.

Given that the great majority of the people in the world commonly use full stops and decimal numbers, the proposed code points would enjoy a world-wide user base.

The distinction between the two meanings of the period is not novel, as Unicode already does so in Arabic block (where the visual difference between the two characters may be more evident):

- ARABIC FULL STOP (U+06D4), '۔'
- ARABIC DECIMAL SEPARATOR (U+066B), '٫'

NB: To those that may not know: the numerals used in the middle East are different from ours. Since our numerals are sometimes called Arabic numerals, theirs may be called Eastern-Arabic numerals to convey the distinction.

Why this is useful

If these new code points are added to Unicode and become actively used in place of the obsolete ASCII code point, it would lead to the following advantages:

1. Computers will be able to distinguish the meaning of the character unambiguously,
2. Users could be shown different graphemes that may be more appropriate for the context or

the locale. For example, the full stop might have some extra space added after it (just like professional word processors like LaTeX already do). But more importantly, the decimal separator could be shown as a period to some users and as a comma to other users, according to what is appropriate in their locale. This is an important thing, so let me elaborate further.

Decimal separators in the world

The world is divided on whether the decimal separator should be a period or a comma. The use of a period is prevalent in countries that have been part of the British commonwealth, and yet some more (like China). The comma, instead, is generally used in almost all other countries (with the exception of some Arabic countries, Iran and Afghanistan). This is shown very clearly in Figure 1.
The diffusion of computers and American culture all over the world has the effect to change the habits of some categories of people, such as computer programmers and scientists, into using periods even in countries where this is not customary, but this fact is irrelevant and cannot be used as an excuse to force the use of periods to other cultures.
Has it ever happened to you to try to load a CSV file is Excel and having it fail because of the wrong locale settings? Or to see a computer program that requires to use commas in one place and periods in others? If not, then you most probably live in a country that does not use commas as decimal separators. To all the others, this leads to frustrations and even more ambiguity.
The dedicated decimal separator proposed here can help this situation. Being appropriately recognized by the software as a decimal separator, it can be displayed in accordance to the locale of the surrounding text: as a period in English text, as a comma in Italian text, and so on. Therefore, on the very same computer, you could read "It costs $1.20" and "Costa $1,20", even though the decimal separator has the same Unicode code point, and it would change from one representation to the other as you copy-paste it from the English text to the Italian text and vice-versa.
The use of locale-dependent representations is not a novelty either. Since a long time, for example, Microsoft Word automatically changes your quote signs into what is appropriate for the locale in which you are writing (see the "Quotation mark" page on Wikipedia for more information about the different quoting styles in the world). Admittedly, in this particular case, Word is transparently replacing the code points. But even then, there are already cases where the very same Unicode code

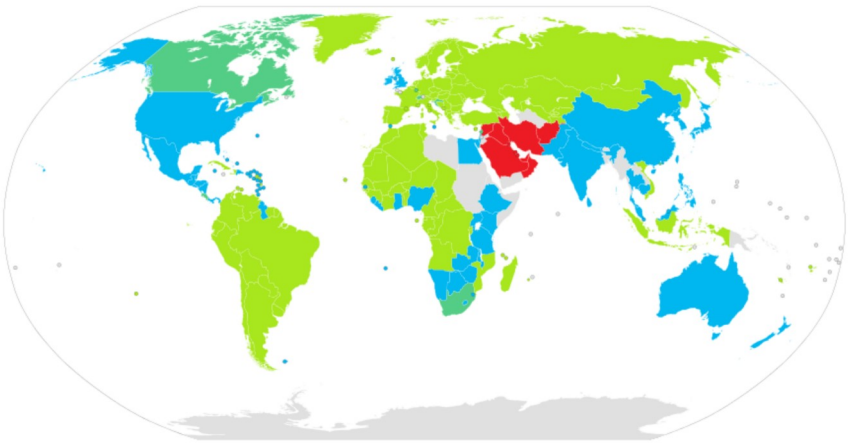

Figure 1: The decimal point in the world:
- countries that use a period '.' are in blue,
- countries that use a comma ',' are in green,
- countries that use the Arabic decimal separator are in red.

point is visualized differently in different locales. For example, in Rumenia the 's' and 't' with cedilla ('ş' and 'ţ') are visualized as 's' and 't' with an underlying comma ('ș' and 'ț'), because this is the current custom in that country. This work is often done automatically by the software libraries used for text rendering, like Pango and Harfbuzz, so that all software based on those libraries behave correctly even out of the knowledge of their programmers.
Finally, one may wonder whether DECIMAL SEPARATOR should be displayed as an Arabic decimal separator in the Arabic locale. My opinion is that it shouldn't, because the Arabic locale already has digits with different graphemes and different code points, and already has a dedicated code point for the decimal separator, too. The two sets should not be mixed: one should use ARABIC DECIMAL SEPARATOR (U+066B) between Eastern-Arabic digits (٠١٢٣٤٫٥٦٧٨٩), but DECIMAL SEPARATOR between "our" Arabic digits (01234.56789).

Other characters that may be suitable for inclusion

Until now, I kept focused on the necessity to distinguish the decimal separator from the other uses of the period. However, even the orthographic full stop is sometimes used for other things than marking the end of the sentence. For example, it is used for abbreviations and acronyms, such as "prof." and "U.S.A." Therefore, it could make sense to add at least one more code point, ABBREVIATION SIGN. Word processors would then be free to treat these periods differently than full stops. Abbreviating periods, for example, do not need any whitespace after them, so word processors should not add whitespace against the expectations of the user.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**
See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** *Two distinct code points: DECIMAL SEPARATOR and FULL STOP*
2. Requester's name: *Dario Schiavon*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *2017-09-08*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *X*
   (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): *No*
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *Yes*
      Name of the existing block: *General Punctuation or Supplemental Punctuation*
2. Number of characters in proposal: *2*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary X | B.1-Specialized (small collection) | B.2-Specialized (large collection)
   C-Major extinct | D-Attested extinct | E-Minor extinct
   F-Archaic Hieroglyphic or Ideographic | G-Obscure or questionable usage symbols
4. Is a repertoire including character names provided? *X*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *annex unavailable*
   b. Are the character shapes attached in a legible form suitable for review? *Unnecessary*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   Unnecessary, no new glyphs are introduced
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
   Unnecessary
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *No*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *Unnecessary*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *Yes*
   It addresses locale-dependent presentation of DECIMAL SEPARATOR.

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *No*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *No*
   If YES, with whom?
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *Yes*
   Reference: *"the proposed code points would enjoy a world-wide user base"*
4. The context of use for the proposed characters (type of use; common or rare)
   Reference: *"commonly"*
5. Are the proposed characters in current use by the user community? *Yes*
   If YES, where? Reference: *"the great majority of the people in the world commonly use full stops and decimal numbers"*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *Yes*
   If YES, is a rationale provided? *Yes*
   If YES, reference: *"To promote the use of these code points, they definitely need to be in the basic multilingual plane"*
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *No*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *Yes*
   If YES, is a rationale for its inclusion provided? *Yes*
   If YES, reference: *"it would lead to the following advantages [...]"*
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *Yes*
   If YES, is a rationale for its inclusion provided? *Yes*
   If YES, reference: *"it would lead to the following advantages [...]"*
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? *Yes*
    If YES, is a rationale for its inclusion provided? *Yes*
    If YES, reference: *"it would lead to the following advantages [...]"*
11. Does the proposal include use of combining characters and/or use of composite sequences? *No*
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *No*
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *No*
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility characters? *No*
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: