L2/24-182R

Unicode request for ExtIPA cartouche

Kirk Miller, kirkmiller, gmail.com
Martin Ball, m.j.ball, bangor.ac.uk

2024 September 04

This proposal, supported by the International Clinical Phonetics and Linguistics Association (p.c. 2024: Sharynne McLeod, president; Joanne Cleland, vice-president), requests two characters to support a transcription convention in the Extensions to the IPA for Disordered Speech (extIPA) that the 1999 IPA *Handbook* calls a 'balloon' and that we will call a 'cartouche.' The symbol chart for the latest revision to the extIPA, which illustrates the cartouche, is publicly available online on the website of the International Phonetic Association.

Problem

The 2016 edition of the extIPA chart marks transcriptions of unidentifiable or indeterminate sounds by enclosing them in a circle (figure 1). The 2008 and earlier editions (figures 2–4) used a typewriter hack: "parentheses linked by a superscript and a subscript line, (‾‾ ‾‾), to enclose the transcription. This is a typescript version of the handwritten balloon" (Duckworth *et al.* 1990: 278).

This cartouche was not requested in L2/20-039 *Unicode request for extIPA support* because we hoped that it could be handled with U+20DD COMBINING ENCLOSING CIRCLE, ◌⃝, with wild-card letters in place of the strings seen in early charts [*e.g.* P⃝ for earlier (P̲l̲.v̲l̲s̲)] and with the letters of extended transcription individually encircled, as U+20DD should contain only a single preceding character, while in manuscript the oval may enclose a longer sequence. According to John Hudson at Tiro Works, the problem with using U+20DD even for a single character is that it would be messy and difficult to adjust the advance width of the character that it encloses, and that any such adjustment would be limited by the scope of the mark-to-base coverage defined by Unicode, with the consequence that for most fonts manual spacing or kerning would be needed to prevent the enclosing circle from overstriking the preceding character. It would be a non-starter for enclosing multiple base characters.

An objection raised to a draft of this proposal was that U+20DD, like any character, can adjust the advance widths of other characters using contextual positioning. However, this is impractical. Victor Gaultney, the font designer at SIL who handles their IPA fonts, explained that the complexity of the required OpenType code is "not worth the effort for font designers to implement," and that in his estimation font designers will not implement it. He outlined the following procedure that would allow U+20DD to enclose a single IPA character, noting that "alternative techniques have similar complexity":

- measure the size and shape of the preceding character, for example by checking if it is in a particular width class;
- turn U+20DD into the needed size and shape, which could require from 20 to 200 extra glyphs due to ascenders, descenders, etc.;
- kern the IPA glyph with the one that precedes it, and also accommodate the situation where it occurs at the beginning of a text run.

Gaultney notes that SIL does something like this for encircling forms in Burmese script, but that the number of possible combinations in Burmese is much more limited. It is also required for Burmese script, but U+20DD is not required for Latin, and this level of complexity makes it impractical.

Proposed solution

Per Gaultney's advice, we propose instead two new Unicode characters to cap the left and right ends of the enclosed string. These would define the beginning and end of scope for layout, both at a plain-text level and as a trigger to additional font-level options to display a proper cartouche. The result would be legible without special font support, and would work with transcriptions of indefinite length. (Although long strings are no longer illustrated in the more aesthetically typeset 2016 version of the extIPA chart, they remain normal practice.)

A pair of bracketing end caps would allow any single letter to be enclosed:

Figure a. ⟮ + x + ⟯ ⇒ ⓧ.

They would allow limited diacritics to be used on the enclosed letter:

Figure b. ⟮ + x + ◌̃ + ◌́ + ◌̥ + ⟯ ⇒ ⓧ̥̃́.

If the enclosed string were longer, a gap would appear between the end caps in a font that did not have the ability to link them. Nonetheless, we believe that this is the best general approach:

Figure c′. ⟮ + x + ʸ + ⟯ ⇒ ⓧʸ,
⟮ + x + y + ⟯ ⇒ ⓧy,
⟮ + x + y + z + ⟯ ⇒ ⓧyz.

With proper font support, figure c′ would display like this:

Figure c″. ⟮ + x + ʸ + ⟯ ⇒ ⓧʸ,
⟮ + x + y + ⟯ ⇒ ⓧy,
⟮ + x + y + z + ⟯ ⇒ ⓧyz.

Without proper font support, linking elements might be added to fill in any gaps:

Figure d. ⟮ + x + ◌͡ + ◌͟ + y + ⟯ ⇒ ⓧy,
⟮ + x + ◌͡ + ◌͟ + y + ◌͡ + ◌͟ + z + ⟯ ⇒ ⓧyz.

We believe that option (d) is not a good solution in general, because it interrupts the transcription with formatting marks that do not need to be preserved in the encoding. Ideally, the font would automatically join up the end caps in figure (c′) to produce figure (c″). This is currently difficult with Open Type fonts, but leaving gaps as seen in figure (c′) would be perfectly intelligible.

Researchers recording their data electronically would presumably choose to use the end caps only, as in figures (c), but for aesthetics in typesetting a publisher might choose to add connecting elements as in figure (d).

For IPA usage, the cartouche would need to enclose both Latin and Greek script without breaking into different shaping runs. If it were extended to Cyrillic phonetic notation, it would need to handle all three scripts. In Windows and presumably other platforms, these scripts use the same rendering engine, so that should not be a problem, though it might create more work for a supporting Open Type font.

We considered other options. When we asked Andrew Glass about the format controls of the Egyptian cartouche as a model for how we might encircle extIPA strings, he said (p.c. 2021) that he thought that the Egyptian implementation would be both too much and insufficiently flexible: *I would recommend encoding a pair of dedicated IPA encircling end caps that, in a suitable font, have the effect of automatically extending over enclosed text in Latin script. ... [Extention over the enclosed text] would be a font choice. Similarly, a font might adapt if a diacritic were present. The key to me is that this special behaviour would be better to associate with new characters rather than existing characters in order to avoid unintended results in a font that did support the adaptive behaviour (diacritic height adjustment or extension). Generally speaking, I think it would be good to aim for the preferred typographic effect even if a subset of fonts support it.*

Ken Whistler agreed: *Overloading the semantics and formatting for some existing pair of common-use parenthesis-type symbols would also not be a good idea. ... And having a dedicated pair of these Latin cartouche end caps would have a viable fallback for fonts that didn't support the full behavior.*

Victor Gaultney, who decided against implementing U+20DD for extIPA use because of the problems noted above, said (p.c.), *if Unicode would approve a Latin cartouche pair that would be an excellent way forward. We could support an unconnected rendering right away, and consider a mechanism to connect them sometime in the future*.

SEW response

The response from SEW, summarized in L2/24-166, was that the extIPA cartouche should not be handled as plain text. Rather, it is equivalent to a copy-editor's mark and should be handled with the same kind of markup. One of us (Martin Ball) is now checking with the rest of the ICPLA on the appropriateness of repurposing an existing pair of Unicode brackets, such as ⸢...⸣ or ⦗...⦘, as a print substitute for the cartouche.

Characters

(U+2E5E LEFT CARTOUCHE END CAP.

) U+2E5F RIGHT CARTOUCHE END CAP.

Properties

2E5E;LEFT CARTOUCHE END CAP;Ps;0;ON;;;;;Y;;;;;
2E5F;RIGHT CARTOUCHE END CAP;Pe;0;ON;;;;;Y;;;;;

Bidi values

The end caps have the bidi-mirrored property “Yes”. The following are the bidi-mirroring glyph values for BidiMirroring.txt:

2E5E; 2E5F # LEFT CARTOUCHE END CAP
2E5F; 2E5E # RIGHT CARTOUCHE END CAP

and for BidiBrackets.txt:

2E5E; 2E5F; o # LEFT CARTOUCHE END CAP
2E5F; 2E5E; c # RIGHT CARTOUCHE END CAP

Linebreaking values

The end caps have the following linebreaking properties for LineBreak.txt:

2E5E; OP # Ps LEFT CARTOUCHE END CAP
2E5F; CP # Pe RIGHT CARTOUCHE END CAP

Annotations

The extIPA cartouch is similar in function to the Egyptian cartouche, but different in its implementation.

2E5E LEFT CARTOUCHE END CAP
→ 13779 EGYPTIAN HIEROGLYPH V011A
2E5F RIGHT CARTOUCHE END CAP
→ 1377B EGYPTIAN HIEROGLYPH V011C

Chart

The end caps should be treated as punctuation marks, analogous to parentheses, rather than as combining marks. The combining behaviour needs to be handled by the font.

Supplemental Punctuation

2E00 **2E7F**

	2E0	2E1	2E2	2E3	2E4	2E5	2E6	2E7
0								
1								
2								
3								
4								
5								
6								
7								
8								
9								
A								
B								
C								
D								
E								
F								

References

Martin Ball, Sara Howard & Kirk Miller (2018) 'Revisions to the extIPA chart', *Journal of the International Phonetic Association*, volume 48, issue 2, pp. 155–164, doi: 10.1017/S0025100317000147. Published online by Cambridge University Press, 11 April 2017.

Chart available on the website of the International Phonetic Association: www.internationalphoneticassociation.org/sites/default/files/extIPA_2016.pdf

Martin Ball, Nicole Müller & Ben Rutter (2010) *Phonology for Communication Disorders*. Taylor & Francis.

Martin Duckworth, George Allen, William Hardcastle & Martin Ball (1990), Extensions to the Internaational Phonetic Alphabet for the transcription of atypical speech. *Clinical Linguistics & Phonetics*, 4 (4) 273–280.

Sara Howard & Anette Lohmander, eds. (2011) *Cleft Palate Speech: Assessment and Intervention*. Wiley-Blackwell.

Barry Hesselwood & Sara Howard (2008) 'Clinical Phonetic Transcription.' In Ball, Perkins, Müller & Howard (eds.) *The handbook of Clinical Linguistics*. Blackwell.

Figures

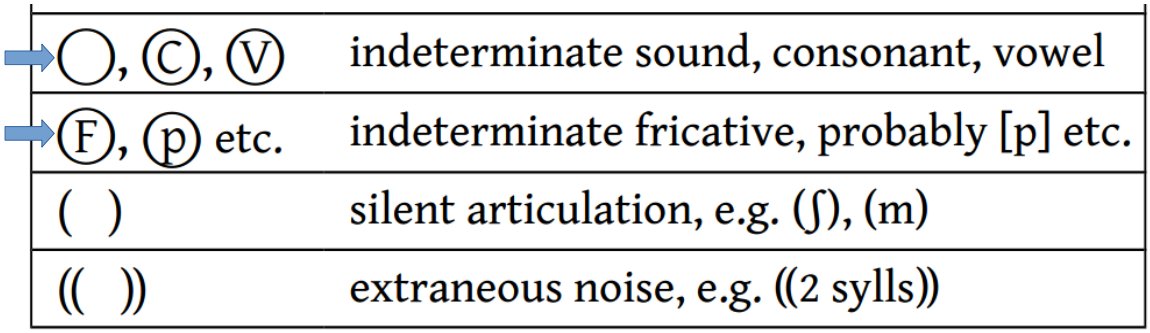

Figure 1. Ball *et al.* (2018: 160). The major 2016 revision of the chart, with single letters circled. This is the subchart for 'connected speech, uncertainty, etc.' The full chart is available at: internationalphoneticassociation.org/sites/default/files/extIPA_2016.pdf.

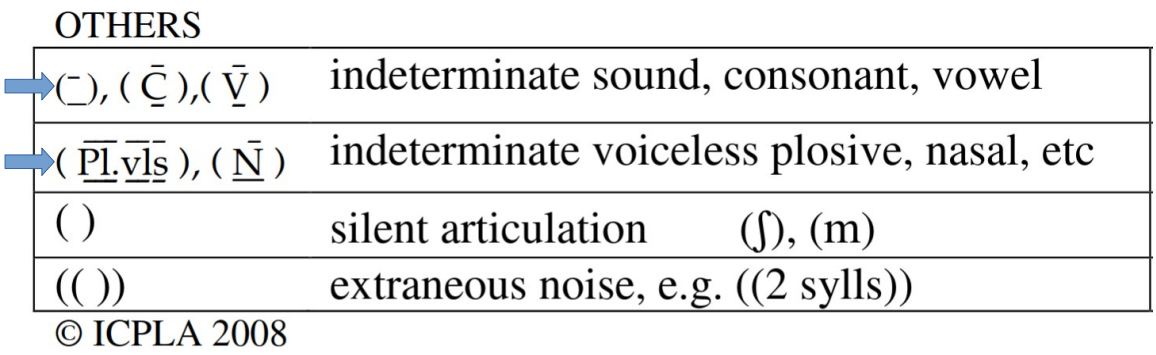

Figure 2. Ball *et al.* (2010: 252). The 'others' section of the 2008 edition of the chart, with letters and longer strings encircled via a typewriter hack. With the proposed end caps, these would be typeset ◯, Ⓒ, Ⓥ, Ⓟl.vl̄s̄, Ⓝ. In a font that properly linked the end caps, the *Pl.vls* string would display as Ⓟl.vls.

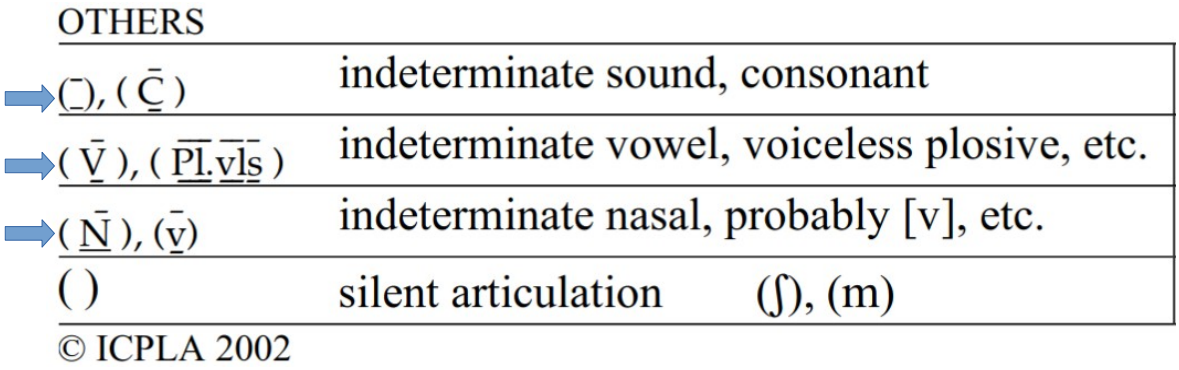

Figure 3. Howard & Lohmander (2011). The 'others' section of the 2002 edition of the chart.

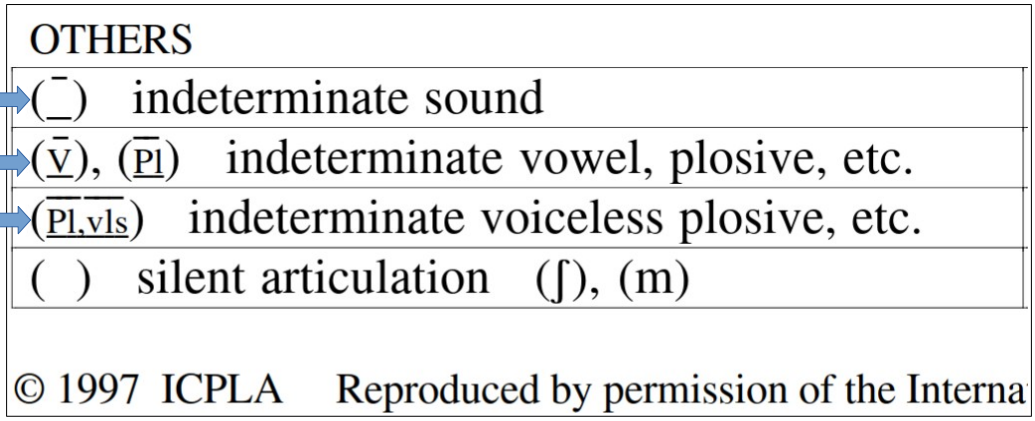

Figure 4. The 'others' section of the 1997 edition of the chart. Chart available at www.arts.gla.ac.uk/IPA/ExtIPAChart97.pdf.

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:** *ExtIPA cartouche*
2. Requester's name: *Kirk Miller, Martin Ball*
3. Requester type (Member body/Liaison/Individual contribution): *individual*
4. Submission date: *2024 September 04*
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):
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *x*
      Name of the existing block: *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? *yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? yes
   b. Are the character shapes attached in a legible form suitable for review? *yes*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard? *Kirk Miller*
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.): *SIL (Gentium release)*
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *yes*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *yes*
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*

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.)? *yes*
   If YES, with whom? *International Clinical Phonetics and Linguistics Association. Proposers are members of the user community.*
   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?
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *phonetic*
   Reference:
5. Are the proposed characters in current use by the user community? *yes*
   If YES, where? Reference: *see References section*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *preferred*
   If YES, is a rationale provided?
   If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *yes*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *no*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *no*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
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?
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences? *yes*
    If YES, is a rationale for such use provided? *yes*
    If YES, reference: *(see refs)*
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    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: