

Unicode request for ExtIPA cartouche

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

2024 February 20

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, \odot , with wild-card letters in place of the strings seen in early charts [e.g. \odot for earlier (P̄.v̄ls)] and with the letters of extended transcription individually encircled. The problem with using U+20DD is that it cannot adjust the advance width of the character that it encloses, with the consequence that without manual spacing or kerning it will overstrike a preceding character. The only non-manual solution would be the impractical one of creating a specialty font with a substitution character for every combination of IPA letter and \odot , with internal anchor points for diacritics that would now need to be input after the circle.

Per the advice of Victor Gaultney at SIL, we propose instead two new Unicode characters to cap the left and right ends of the enclosed string. The result would be legible even 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. $(+ x +) \Rightarrow (\text{x})$.

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

Figure b. $(+ x + \tilde{\circ} + \acute{\circ} + \circ +) \Rightarrow (\text{x})$.

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:

$$\begin{aligned}\text{Figure c'}. \quad & (+ x + y +) \Rightarrow (\overline{x^y}), \\ & (+ x + y +) \Rightarrow (\overline{xy}), \\ & (+ x + y + z +) \Rightarrow (\overline{xyz}).\end{aligned}$$

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

$$\begin{aligned}\text{Figure c''}. \quad & (+ x + y +) \Rightarrow (\overline{x^y}), \\ & (+ x + y +) \Rightarrow (\overline{xy}), \\ & (+ x + y + z +) \Rightarrow (\overline{xyz}).\end{aligned}$$

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

$$\begin{aligned}\text{Figure d}. \quad & (+ x + \overline{\circ} + \underline{\circ} + y +) \Rightarrow (\overline{xy}), \\ & (+ x + \overline{\circ} + \underline{\circ} + y + \overline{\circ} + \underline{\circ} + z +) \Rightarrow (\overline{xyz}).\end{aligned}$$

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).

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 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. ... [Extension 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, the font designer at SIL 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.*

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.

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

	2E00	2E01	2E02	2E03	2E04	2E05	2E06	2E07
0	ⵀ	ⵁ	ⵂ	ⵃ	ⵄ	ⵅ		
1	ⵆ	ⵇ	ⵈ	ⵉ	ⵊ	ⵋ		
2	ⵌ	ⵍ	ⵎ	ⵏ	ⵐ	ⵑ		
3	ⵒ	ⵓ	ⵔ	ⵕ	ⵖ	ⵗ		
4	ⵘ	ⵙ	ⵚ	ⵛ	ⵜ	ⵝ		
5	ⵞ	ⵟ	ⵠ	ⵡ	ⵢ	ⵣ		
6	ⵤ	ⵥ	ⵦ	ⵧ	⵨	⵩		
7	⵪	⵫	⵬	⵭	⵮	ⵯ		
8	⵰	⵱	⵲	⵳	⵴	⵵		
9	⵶	⵷	⵸	⵹	⵺	⵻		
A	⵼	⵽	⵾	⵿	ⶀ	ⶁ		
B	ⶂ	ⶃ	ⶄ	ⶅ	ⶆ	ⶇ		
C	ⶈ	ⶉ	ⶊ	ⶋ	ⶌ	ⶍ		
D	ⶎ	ⶏ	ⶐ	ⶑ	ⶒ	ⶓ		
E	ⶔ	ⶕ	ⶖ	⶗	⶘	⶙		
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 International 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

→ ○, ©, ⑤	indeterminate sound, consonant, vowel
→ ①, ② etc.	indeterminate fricative, probably [p] etc.
()	silent articulation, e.g. (ʃ), (m)
(())	extraneous noise, e.g. ((2 sylls))

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.

OTHERS

→ (□), (\bar{C}), (\bar{V})	indeterminate sound, consonant, vowel
→ ($\bar{Pl.vls}$), (\bar{N})	indeterminate voiceless plosive, nasal, etc
()	silent articulation (ʃ), (m)
(())	extraneous noise, e.g. ((2 sylls))

© ICPLA 2008

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 \bigcirc , \bigcirc , \bigcirc , \bigcirc , \bigcirc . In a font that properly linked the end caps, the *Pl.vls* string would display as \bigcirc .

OTHERS

→ (□), (\bar{C})	indeterminate sound, consonant
→ (\bar{V}), ($\bar{Pl.vls}$)	indeterminate vowel, voiceless plosive, etc.
→ (\bar{N}), (\bar{v})	indeterminate nasal, probably [v], etc.
()	silent articulation (ʃ), (m)

© ICPLA 2002

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

OTHERS	
→ (□)	indeterminate sound
→ (\bar{V}), (\bar{Pl})	indeterminate vowel, plosive, etc.
→ ($\bar{Pl.vls}$)	indeterminate voiceless plosive, etc.
()	silent articulation (ʃ), (m)
© 1997 ICPLA Reproduced by permission of the Interna	

Figure 4. The ‘others’ section of the 1997 edition of the chart. Chart available at www.arts.gla.ac.uk/IPA/ExtIPACChart97.pdf.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹.

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:	<i>ExtIPA cartouche</i>
2. Requester's name:	<i>Kirk Miller, Martin Ball</i>
3. Requester type (Member body/Liaison/Individual contribution):	<i>individual</i>
4. Submission date:	<i>2024 February 20</i>
5. Requester's reference (if applicable):	
6. Choose one of the following:	
This is a complete proposal:	<i>x</i>
(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:	<i>x</i>		
Name of the existing block:	<i>Supplemental Punctuation</i>		
2. Number of characters in proposal:	<i>2</i>		
3. Proposed category (select one from below - see section 2.2 of P&P document):			
A-Contemporary <i>x</i>	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?			
a. If YES, are the names in accordance with the “character naming guidelines” in Annex L of P&P document?	<i>yes</i>		
b. Are the character shapes attached in a legible form suitable for review?	<i>yes</i>		
5. Fonts related:			
a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?	<i>Kirk Miller</i>		
b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):	<i>SIL (Gentium release)</i>		
6. References:			
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?	<i>yes</i>		
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?	<i>yes</i>		
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)?			<i>yes</i>

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? <i>International Clinical Phonetics and Linguistics Association.</i>	
<i>Proposers are members of the user community.</i>	
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: <i>see References section</i>	
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?	
If YES, reference: <i>(see refs)</i>	
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:	