

Unicode request for Harrington diacritics

Yoram Meroz, yoram.meroz, gmail
Matthew Vestuto, mvestuto, gmail

Kirk Miller, kirkmiller, gmail
2023 September 20

This is a request for combining diacritics found in the manuscripts of John Peabody Harrington and used by linguists and archivists transcribing his work.

For many languages of the US Pacific Coast, Harrington’s materials are the most extensive and reliable documentation available. Increasing use of his material is being made by indigenous communities for language-revival projects. In proposal L2/20-182, one of us requested a combining superscript number sign, U+1AC6 [♯], for use in a Chumash dictionary project that relied on Harrington’s work. It turns out that the diacritic is actually a musical sharp sign for high pitch: <[♯]>, and is paired with a combining flat sign for low pitch: <[♭]>. Lisa Woodward, archivist at the Pechanga Cultural Resources Department (Luiseño) and formerly at the Harrington Database Project at UC Davis, notes that the combining flat and sharp signs are quite common in the Uto-Aztecan languages that she’s familiar with (p.c. 2023).

We request that both be adopted, and three additional diacritics:

A combining down tack–above ⟨◌̣̤⟩ is found in Harrington material as a middle-pitch diacritic (Sandy 2024: xi), and coincidentally is also needed for IPA, where it may substitute for U+031E COMBINING DOWN TACK BELOW ⟨◌̥̦⟩ on letters that have descenders. Sandy notes that in Harrington’s Karuk material ⟨◌̣̤⟩ contrasts with acute and grave accents for high and low pitch. (The macron is used for vowel length.) The remainder of Harrington’s Karuk diacritics already have Unicode support.

A non-stacking dot and ring, $\langle \circ \rangle$, is also requested. The SAH has noted in the past that it is inadvisable to require special stacking rules for diacritics, so pairs of adjacent diacritics such as $\langle \circ \rangle$ should be encoded individually.

The last diacritic is two dots orientated like a diaeresis rotated to a grave angle: <̣̣>. This might be substituted with U+0307 COMBINING DOT ABOVE plus U+1DF8 COMBINING DOT ABOVE LEFT; however, U+1DF8 is specified as being for Syriac and Typicon Cyrillic, so font support may be a problem. We suggest the name COMBINING FALLING DIAGONAL DIAERESIS in order to disambiguate a similar, but acute-angled, MUFI diacritic, which if accepted into Unicode could be named COMBINING RISING DIAGONAL DIAERESIS.

Finally, we identify Harrington’s nesting carets as the standard copyedit marks for raising and lowering a character off the baseline (Figures 17–20). That is, they are a manuscript convention for marking superscript and subscript. The SAH advises that they should not be encoded by Unicode unless someone wishes to propose copyedit symbols more generally and on their own merit. We illustrate them here for reference.

Dedicated sharp-sign diacritic

It is debatable whether the proposed combining sharp sign and the existing combining number sign U+1AC6 should be unified or disunified. If unified, U+1AC6 would need its annotation changed from “superscript octothorp” to “superscript music sharp sign,” and a name alias of the type “correction” with the value “COMBINING SHARP SIGN.” Its glyph would also need to be changed to a music sharp sign: ♯. It would moreover be problematic to unify them when the base characters are distinguished in Unicode.

If disunified, U+1AC6 would be left as an orphan, at least initially. However, we have found a true combining octothorp, which U+1AC6 might eventually be used for, in the Stokoe notation of Italian Sign Language (Figure 1).



Figure 1. Unicode proposal L2/12-133, fig 16, from Radutzky (1992: 32). The ‘5’ hand shape with a combining ‘closed’ diacritic (♯), followed by modifier diacritics for ‘motion toward the signer’ and ‘motion upward’. Although Stokoe notation is not (yet) supported by Unicode, if it ever is, U+1AC6 could be used for this diacritic.

Adopting a new code point for the Harrington combining sharp sign should not be a problem for existing encoding. One of us (Matthew Vestuto) is the Chairman of the Barbareño/Ventureño Band of Mission Indians (Chumash), its language program coordinator, director of the Tšumaš Transcription Project, and is on the board of Advocates for Indigenous California Language Survival. Thus far, none of those entities utilize U+1AC6 anywhere but in raw transcripts. It will be easy to change U+1AC6 to a new code point with find/replace. Timothy Henry-Rodriguez, co-author of Unicode request L2/20-182 for U+1AC6, and a member of the Purisimeño Chumash dictionary project that had originally needed the diacritic, says that he knows of no other electronic sources that use U+1AC6 or that would be adversely impacted by Unicode adopting a new sharp-sign diacritic (p.c. 2023).

Chart

The characters in white cells are proposed in this request. Assigned cells are medium grey, apart from the combining number sign in pink. Light-grey cells are proposed in separate requests.

Combining Diacritical Marks Extended

1AB0

1AEF

	1AB	1AC	1AD	1AE
0	◌̂	◌̇	◌̈	
1	◌̉	◌̊	◌̋	
2	◌̌	◌̍	◌̎	
3	◌̏	◌̐	◌̑	
4	◌̒	◌̓	◌̔	
5	◌̖	◌̗	◌̘	
6	◌̙	◌̚	◌̛	
7	◌̜	◌̝	◌̞	
8	◌̟	◌̠	◌̡	
9	◌̢	◌̣	◌̤	
A	◌̥	◌̦	◌̧	
B	◌̨	◌̩	◌̪	
C	◌̬	◌̭	◌̮	
D	◌̯	◌̰	◌̱	
E	◌̲	◌̳		
F	◌̴	◌̵		

Characters

- # U+1AD9 COMBINING SHARP SIGN. Figures 2, 12.
- ◌ U+1ADA COMBINING FLAT SIGN. Figure 2.
- ◌ U+1ADB COMBINING DOWN TACK ABOVE. Figures 7–16.
- ◌ U+1ADC COMBINING FALLING DIAGONAL DIAERESIS. Figure 6.
- ◌ U+1ADD COMBINING DOT-AND-RING BELOW. Figures 3–5.

Properties

1AD9;COMBINING SHARP SIGN;Mn;230;NSM;;;;;N;;;;;
1ADA;COMBINING FLAT SIGN;Mn;230;NSM;;;;;N;;;;;
1ADB;COMBINING DOWN TACK ABOVE;Mn;230;NSM;;;;;N;;;;;
→ U+031E COMBINING DOWN TACK BELOW
1ADC;COMBINING FALLING DIAGONAL DIAERESIS;Mn;230;NSM;;;;;N;;;;;
1ADD;COMBINING DOT-AND-RING BELOW;Mn;220;NSM;;;;;N;;;;;

References

- John P. Harrington Papers, digitized microfilm edition. Smithsonian Institution.
- NY Book Editors (revised from 2013). *A Guide to Copyediting Marks*. Available online at <https://nybookeditors.com/2013/06/copyediting-marks/>.
- Radoslav Pavlík (2004) Slovak Speech Sounds and the International Phonetic Alphabet. *Jazykovedný časopis*, 55, No. 2, pp. 87–109 (Bratislava).
- Real Academia Española (2009–2011) *Nuevo gramática de la lengua española*. Asociación de academias de la lengua española.
- Clare Sandy (to appear 2024) Recovering Prosody from Karuk texts: Deciphering J.P. Harrington’s diacritics. In Darya Kavitskaya & Alan C. L. Yu (eds.), *The life cycle of language: Past, present, and futur*. Oxford University Press.
- Suzanne Wash (2001) *Adverbial Clauses in Barbareño Chumash Narrative Discourse*. PhD dissertation, University of California Santa Barbara.

Figures

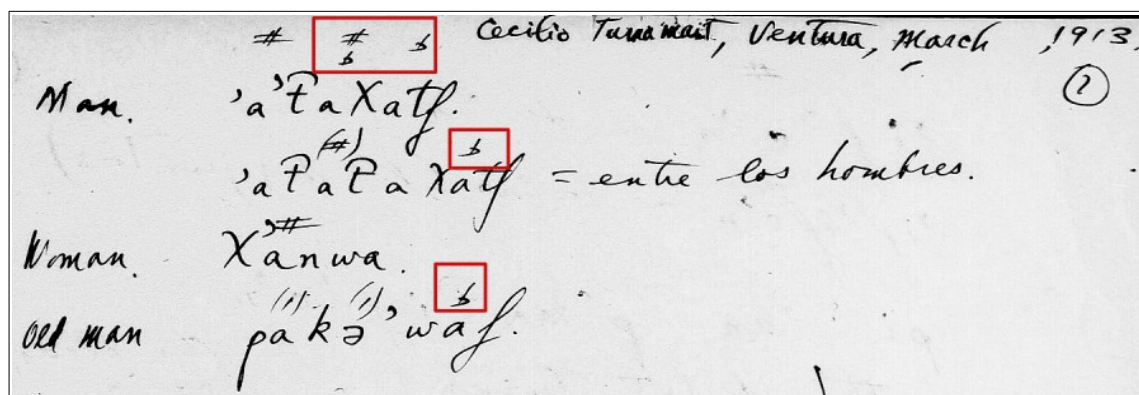


Figure 2. Harrington (series 3, reel 69, frame 24). The sharp and flat signs <#, b> for pitch. We believe the stacked flat and sharp signs indicate falling intonation.

o	X
o	X
o	X

Figure 3. Combinations of dot and ring <.,> in the list of Harrington symbols kept by the Harrington Database Project.

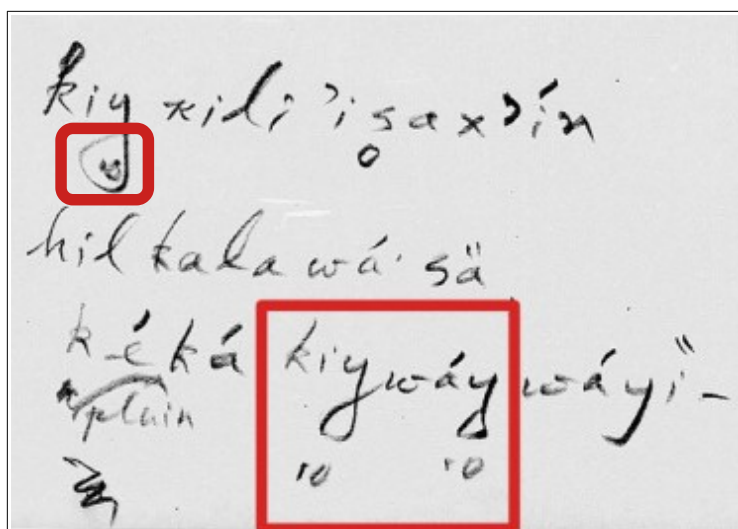


Figure 4. Harrington (series 3, reel 59, frame 645). Manuscript transcription of Barbareño Chumash, with the dot on the left and the ring on the right.

	<u>Harrington:</u>	<u>This dissertation:</u>
uvular stop	κ	κ
double accents:	â, ê, î, ô, û	a ^{˘˘} , e ^{˘˘} , i ^{˘˘} , o ^{˘˘} , u ^{˘˘}
circumflex:	â, ê, î, ô, û	a [˘] , e [˘] , i [˘] , o [˘] , u [˘]
circumflex with macron:	ā, ē, ī, ō, ū	a ^{˘ˉ} , e ^{˘ˉ} , i ^{˘ˉ} , o ^{˘ˉ} , u ^{˘ˉ}
over-ring:	â, ê, î, ô, û	a [˙] , e [˙] , i [˙] , o [˙] , u [˙]
under-ring:	â, ê, î, ô, û	a _˙ , e _˙ , i _˙ , o _˙ , u _˙
under-ring w/dot:	â, ê, î, ô, û	a _{˙˙} , e _{˙˙} , i _{˙˙} , o _{˙˙} , u _{˙˙}
glottalization:	ʔ, ɬ, ʈ, ʈs, etc.	ʔ, ɬ, ʈ, ʈs, etc.
aspiration:	h, pʰ, tʰ, sʰ, etc.	h, pʰ, tʰ, sʰ, etc.
subscript arch:	h̲, hi̲	h _ˉ , h _ˉ i _ˉ

Figure 5. Wash (2001: 27). Wash apparently got the order of dot and ring reversed; we can only attest to a dot followed by a ring in Harrington's material.

Fr. J. C. fra am
 call ghost ^{in V. alt} sinax kuhii =
 Sp. espanto. o espiritu
 ente means that
 which scares.
 tsiwin loka 'enhefif.
 hocen

V. 'al ka'ja'ji =

Figure 6. Harrington (series 3, reel 69, frame 241). The grave-angled diaeresis <◌◌̈> in Ventureño. In manuscript form, this does not replace the tittle of an *i*, perhaps to avoid confusion with the dotless letter *ı* or *l*. Documentarians have typeset these as *ad hoc* <◌̈, ◌̊, ◌̋, ◌̌>, but have expressed a desire for explicit Unicode support.

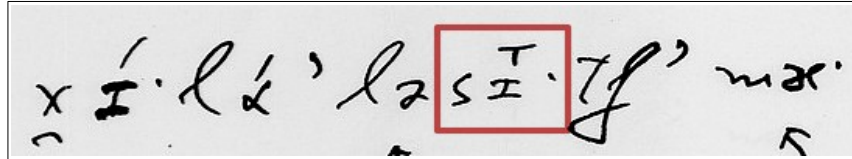
TABLE 0.3 Most common diacritics used by Harrington.		
Diacritic	Example	JPH label
Acute	ạ́	high pitch
Downtack	ạ̄	middle pitch
Grave	ạ̀	low pitch
Acute-tilde	ạ̃	high pitch, falling tone
Inverted breve	ậ	low falling tone
Acute-inverted breve below length	ạ̣́	high pitch, inlaut form of low falling tone
Circumflex below length	ậ̣	inlaut form of low falling atonic
Apostrophe with echo vowel	a'a	[indicates glottal interruption of vowel; not listed as tone.]

Figure 7. Sandy (2024: xi) for Karuk. Down tack <◌̣>, the single unsupported diacritic in Harrington’s Karuk material.

(e) Xó[̣]xhí[̣]rāk ‘Martins Ferry’ (Phoebe Maddux, KIM-08:6)
xôoxhí[̣]rāk ‘Martin’s Ferry’ (placename)

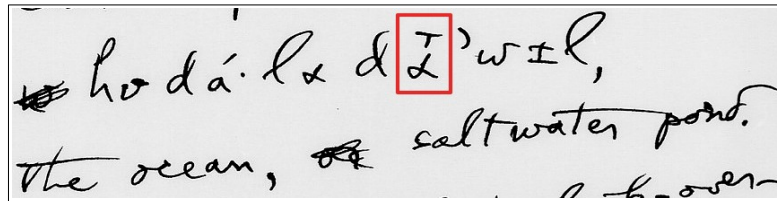
(b) Yuptcúk[̣]kinā[̣]nàtc múttca’s xákka[̣]n kunʔí[̣]n ’Akví’ccitc.

Figure 8. Sandy (2023: xii, xvi). The down tack <̣> replacing the title of an *i*.



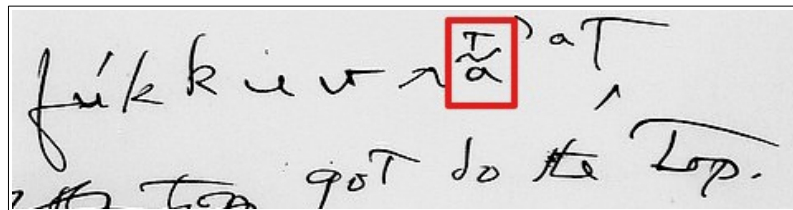
The image shows a handwritten transcription of the word 'Coos'. A red box highlights a superscript down tack (̣) placed over the letter 'i' in the sequence 'si'. The full transcription appears to be 'xí·l'á' l'á s[̣]i·f' mæ'.

Figure 9. Harrington (series 1, reel 24, frame 20). Use of superscript down tack <̣> in the transcription of Coos.



The image shows a handwritten transcription of the word 'Wiyot'. A red box highlights a superscript down tack (̣) placed over the letter 'i' in the sequence 'd'á·l'á d'á[̣]wí·l'. The full transcription appears to be 'hó d'á·l'á d'á[̣]wí·l', with the English translation 'the ocean, saltwater pond' written below.

Figure 10. Harrington (series 2, reel 1, frame 335) for Wiyot.



The image shows a handwritten transcription of the word 'Karuk'. A red box highlights a stacked tilde and down tack (̣̃) placed over the letter 'a' in the sequence 'fú·k k i u r a^{̣̃}at'. The full transcription appears to be 'fú·k k i u r a^{̣̃}at', with the English translation 'the tea got to the top' written below.

Figure 11. Harrington (series 2, reel 12, frame 275). Stacking of a tilde and down tack in the transcription of Karuk.

τ	τ̣	υ
	χ	υτ
#	#̣	υ#

Figure 12. The down tack <̣> and sharp sign <̣̃> in the list of Harrington symbols kept by the Harrington Database Project.

gradualmente. Los alófonos articulados con menor esfuerzo, como la segunda consonante de las palabras *dado* ['daðo] y *gago* ['ga[̞]o] reciben la denominación

segmentos /b/, /d/, /g/ posee dos tipos de alófonos: uno oclusivo —[b], [d] y [g]— y otro aproximante —[β], [ð] y [[̞]]

Figure 13. Real Academia Española (2009–2011: 6278–6279 [§4.2f]). The down tack is below the <ð> in *dado*: <ð>, but above the <y> in *gago*: <y>. This convention is used throughout the work.

UNIDADES A LAS QUE AFECTA Y POSICIÓN EN LA SÍLABA	REALIZACIONES O ALÓFONOS RESULTANTES
/b/ explosiva (especialmente entre vocales) y también implosiva	[β] ~ [β̞] (en función del grado de debilitamiento)
/d/ explosiva (especialmente entre vocales) y también implosiva	[ð] ~ [ð̞] (en función del grado de debilitamiento)
/g/ explosiva (especialmente entre vocales) y también implosiva	[[̞]] ~ [^{̞̞}] (en función del grado de debilitamiento)

Figure 14. *Ibid.* (p. 6448 [§7.3e]). Down tack below on superscript [β̞] and [ð̞] contrasting with down tack above on superscript [^{̞̞}].

	Bilabiálny	Labio-dentálny	Alveolárny	Postal-veolárny	Palatálny	Velárny	Glotá-ly
Explozívny	p b		t d		c ɟ	k g	
Nazálny	m	ɱ	ɳ ̃		ɲ	ŋ ̃	
Vibranty			ɾ ː				
Jednokmitové hlásky			ɻ				
Frikatívy		f v	s z	ʃ ʒ		x ɣ	ɦ
Afrikáty			t͡s d͡z	t͡ʃ d͡ʒ			
Centrálne aproximanty	w	ʋ			j ̰		

Figure 15. Pavlík (2004: 106, Table 3). A superscript down tack <̰> in the IPA transcription of Slovak, where the base letter has a descender or another diacritic. The authors explain (p 101–102) that [j̰] is basically equivalent to <ũ>, but use of the down tack is more appropriate in their context. *Uptack above* <̰> is also used, but that is already supported by Unicode at U+1DF5.

Slovenská fon. abeceda	p	b	t	d	t'	d'	k	g	m	ɱ	n	ɳ	ɲ	ɸ	r	ɾ	ɻ	f	w
IPA	p	b	t	d	c	ɟ	k	g	m	ɱ	ɳ	̃	ɲ	̃	ɻ	ɾ	ː	f	v

Slovenská fon. abeceda	s	z	š	ž	x	ɣ	h	c	ʒ	č	ʒ	ũ	v	j	ĩ	l	ɭ	ɮ	ɹ
IPA	s	z	ʃ	ʒ	x	ɣ	ɦ	t͡s	d͡z	t͡ʃ	d͡ʒ	w	ʋ	j	̰	ɭ	ɮ	ɮː	ɹ̥

Tab. 4. Konfrontačná tabuľka spoluhláskových znakov slovenskej fonetickej abecedy a IPA

Figure 16. Pavlík (2004: 107, Table 4). Correspondence between the Slovak phonetic alphabet and IPA, with down tack above on three IPA letters.

l'á. d'á. 'Vl, the
betw. Mattot R. +

Figure 17. Harrington (series 2, reel 1, frame 455). Superscript alpha for an epenthetic or reduced sound. Harrington uses the standard copyedit mark for a raised letter, presumably so that the distinction between superscript and in-line notation is not confused in rapid transcription.

Steward, Mar. 22, 39:
"East ... taven do i²⁵
↑ a space
as printed, but
Hew's ms. had³
[Fort Hall Shoshoni word for
east]
"25³ Literally sun up
direction."
[Steward, Julian,
Basin Plateau aboriginal
Socio-political group, BAE
Bul 120,
p. 277 [1925]]

Figure 18. Harrington (series 3, reel 171, frame 784). Here Harrington copies an entry from a BAE bulletin, using the caret to mark the superscript footnote number, as well as to mark an apostrophe (raised comma) for glottal stop transcribed in isolation, demonstrating that these are just the standard copyedit marks.

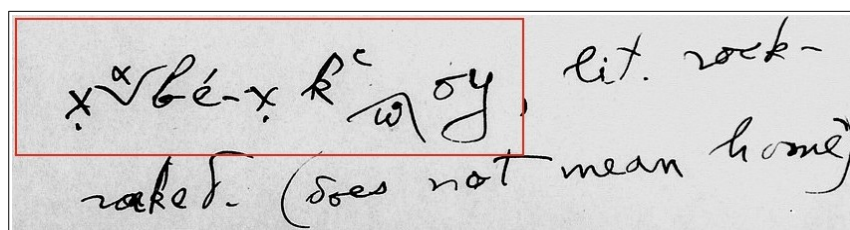


Figure 19. Harrington (series 2, reel 5, frame 304). The corresponding reverse copyedit mark is used to mark subscript w.

	superscript	$E=mc^2$, even if we don't understand it, should be $E=mc^2$.
	subscript	Do we breathe O_2 or O_2 ?
	comma	As you can see, commas get a roof placed over them.
	apostrophe	Its easy to miss these.
	quotation marks	Did I forget I was speaking? she said.

Figure 20. NY Book Editors (revised from 2013). Harrington's usage of the caret is simply the standard copyedit convention for marking superscript and subscript, as well as to clarify punctuation marks such as the apostrophe for glottal stop in Figure 18. In the latter the caret is not an insertion mark – note use of caret-apostrophe together with an insertion caret at right in this figure.

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>Harrington diacritics</i>
2. Requester's name:	<i>Kirk Miller, Yoram Meroz, Matthew Vestuto</i>
3. Requester type (Member body/Liaison/Individual contribution):	<i>individual</i>
4. Submission date:	<i>2023 September 20</i>
5. Requester's reference (if applicable):	
6. Choose one of the following:	
This is a complete proposal:	<i>yes</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):	<i>no</i>
Proposed name of script:	
b. The proposal is for addition of character(s) to an existing block:	<i>yes</i>
Name of the existing block:	<i>Combining Diacritical Marks Extended</i>
2. Number of characters in proposal:	<i>5</i>
3. Proposed category (select one from below - see section 2.2 of P&P document):	
A-Contemporary ☐ 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?	<i>yes</i>
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>no</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? Harrington Database Project; WIELD; BVBMI	
If YES, available relevant documents: (emails available on request)	
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 illustrations	
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?	no
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?	yes
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?	yes
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?	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:	