

# Unicode Characters and Typing Yukon Languages

Doug Hitch, Linguist, Yukon Native Language Centre, Whitehorse Yukon. October 23, 2003

Adobe glyph names from <http://partners.adobe.com/asn/developers/typeforum/unicodegn.html>

Glyph names which are not defined by Adobe are in parentheses. Adobe offers alternates -comb and -cmb for the combining diacritics. -cmb is used here.

Glyphs used in substitutions which are not defined in Unicode as characters are included in these charts in logical places.

Keystrokes are those used in the YNLC Unicode keyboards with Caps Lock on. Macintosh and Windows keystrokes are the same, except that the Macintosh option key is equivalent to the Windows alt-ctrl combination. See the special keyboarding document.

*Basic Latin. 0000–007F.* All included. Plain ASCII.

*Latin-1 Supplement. 0080–00FF.* All included. Mac and Windows higher ascii compatibility. (MS Office requires 00A0 NO-BREAK SPACE, 00A4 CURRENCY SIGN, 00A7 SECTION SIGN, 00AC NOT SIGN, 00B0 DEGREE SIGN, 00B6 PILCROW SIGN, 00B7 MIDDLE DOT.)

*Latin Extended-A. 0100–017F.*

hex	glyph	Unicode name	PostScript name	Macintosh keystrokes	Windows keystrokes	Current Yukon Uses
0131	ı	LATIN SMALL LETTER DOTLESS I	dotlessi	(n/a)	(n/a)	substitutes for dotted i when i has a superscript diacritic. (Turkish)
(0131.1) (Apple and Adobe private use F6BE)	Ĳ	(glyph only)	(dotlessj)	(n/a)	(n/a)	linguistics. consistent with dotlessi
0141	Ł	LATIN CAPITAL L WITH STROKE	Lslash	opt-sh-l	alt-ctrl-sh-l	voiceless lateral: all languages
0142	ł	LATIN SMALL L WITH STROKE	lslash	opt-l	alt-ctrl-l	voiceless lateral: all languages
014A	Ń	LATIN CAPITAL LETTER ENG	Eng	opt-sh-n	alt-ctrl-sh-n	consistent with eng
014B	ń	LATIN SMALL LETTER ENG	eng	opt-n	alt-ctrl-n	linguistics. Inupiaq.
0152	Œ	LATIN CAPITAL LIGATURE OE	OE	opt-sh-o	alt-ctrl-sh-o	consistent with œ. Unknown if in Petitot.
0153	œ	LATIN SMALL LIGATURE OE	oe	opt-o	alt-ctrl-o	Petitot Tukudh

*Latin Ext-B. 0180–024F. No characters*

*IPA Extensions. 0250–02AF*

hex	glyph	Unicode name	PostScript Name	Macintosh keystrokes	Windows keystrokes	Current Yukon Uses
0259	ə	LATIN SMALL LETTER SCHWA	schwa	opt-e	alt-ctrl-e	Yukon linguistics
025B	ɛ	LATIN SMALL LETTER OPEN E	eopen	opt-sh-e	alt-ctrl-sh-e	Yukon linguistics (unclear if epsilon 03B5 better choice)
0294	ʔ	LATIN LETTER GLOTTAL STOP	glottalstop	opt-/	alt-ctrl-/	Yukon linguistics. NWT Athapaskan.

*Spacing Modifier Letters. 02B0–02FF*

hex	glyph	Unicode name	PostScript name	Macintosh keystrokes	Windows keystrokes	Current Yukon Uses
02BC	’	MODIFIER LETTER APOSTROPHE	apostrophemod	’	’	ejective or glottalized consonant: all languages
02EE	”	MODIFIER LETTER DOUBLE APOSTROPHE	(dblapostrophemod)	” or opt-’	”	Petiot Tukudh.
(proposed character)	.	(MODIFIER LETTER PERIOD)	(periodmod)	(n/a)	(n/a)	Tlingit glottal stop

*Combining Diacritical Marks. 0300–036F*. All names begin with COMBINING. Many end with ACCENT. E.g., 0300 is technically COMBINING GRAVE ACCENT. Keystrokes are identical on Macintosh and Windows except that Macintosh has the double acute additionally at option-j.

hex	glyph	Unicode name	PostScript name	key	Current Yukon Uses
0300	̀	GRAVE	gravecmb	] ]	low tone: Gwich'in, Hän, Southern Tutchone, Tagish, Upper Tanana low toned long vowel: Tlingit rising tone: Northern Tutchone
0301	´	ACUTE	acutecmb	[ [	high tone: Northern Tutchone, Kaska, Tlingit rising tone: Southern Tutchone
0302	˘	CIRCUMFLEX	circumflexcmb	{ }	high toned long vowel: Tlingit falling tone: Hän Petitot Tukudh.
0303	˜	TILDE	tildcmb	~	Petitot. Inupiaq, Ummamniutun.
0304	˘	MACRON	macroncmb	:	long vowels: Kaska, Tagish, technical Gwich'in mid-high tone: Northern Tutchone falling tone: Southern Tutchone
(0304.1)	̄	(glyph only) BREVE	(macroncmb) brevecmb	(n/a) `	substitutes for macroncmb above i linguistics, short vowel.
0306	˙	BREVE	brevecmb	`	future use
0307	ˆ	DOT ABOVE	dotaccntcmb	=	
0308	¨	DIAERESIS	dieresiscmb	/	ä as schwa/epsilon: Southern Tutchone i as high central unrounded: Southern Tutchone ä as low back: Hän ë as schwa: Hän Petitot and McDonald Tukudh.
030B	˝	DOUBLE ACUTE	hungarumlautcmb	l (opt-j)	extra high tone: Upper Tanana, Tanacross
030C	˘	CARON hacek	caroncmb	}	rising tone: Hän, Upper Tanana mid tone: Kaska
030F	˘˘	DOUBLE GRAVE	dblgravecmb	\	no current use. corresponds to double acute.
0323	˙	DOT BELOW	dotbelowcmb	sh-period	Petitot Tukudh under s, t, n. McDonald Tukudh under e, i, o, u.
0325	◌̣	RING BELOW	ringbelowcmb	sh-comma	linguistics, voicelessness.
0327	◌̣̣	CEDILLA	cedillacmb	+	French, (Portuguese, Turkish)
0328	◌̣̣̣	OGONEK	ogonekcmb	;	nasalization: all languages
0331	◌̣̣̣̣	MACRON BELOW	macronbelowcmb	–	semi-voiced fricatives: Upper Tanana, Tanacross. goes under t, th, s, sh, x, ʃ
(0331.1)	◌̣̣̣̣̣	(glyph only)	(macronbelowcmb)	(n/a)	substitutes for macronbelowcmb under l, etc.

*Greek and Coptic. 0370–03FF*

hex	glyph	Unicode name	PostScript Names	Macintosh keystrokes	Windows keystrokes	Current Yukon Uses
03C1	ρ	GREEK SMALL LETTER RHO	rho	opt-r	alt-ctrl-r	Petitot Tukudh
03C7	χ	GREEK SMALL LETTER CHI	chi	opt-x	alt-ctrl-x	Petitot Tukudh

*General Punctuation. 2000–206F. Keystrokes unchanged from US.*

2013	–	EN DASH	endash			typography
2014	—	EM DASH	emdash			typography
2018	‘	LEFT SINGLE QUOTATION MARK	quoteleft			MSWord smart quotes, InDesign printer's quotes
2019	’	RIGHT SINGLE QUOTATION MARK	quoteright			MSWord smart quotes, InDesign printer's quotes
201C	“	LEFT DOUBLE QUOTATION MARK	quotedblleft			MSWord smart quotes, InDesign printer's quotes
201D	”	RIGHT DOUBLE QUOTATION MARK	quotedblright			MSWord smart quotes, InDesign printer's quotes
2022	•	BULLET	bullet			MSOffice
2026	…	HORIZONTAL ELLIPSIS	ellipsis			MSWord converts 3 periods to ellipsis

*Mathematical Operators. 2200–22FF. Keystrokes unchanged from US.*

2219	•	BULLET OPERATOR	bulletoperator			MSOffice
------	---	-----------------	----------------	--	--	----------

*Geometric Shapes. 25A0–25FF*

25CC	◦	DOTTED CIRCLE	dottedcircle	opt-0 (zero)	alt-ctrl-0	(used to indicate combining diacritics)
------	---	---------------	--------------	--------------	------------	---

### Unicode Characters Uncoded from the YNLIC Fonts

Alongside the possibility of composing a base letter and a diacritic, Unicode also provides numerous precomposed characters which give a similar visual effect. That is, besides using a plain letter *a* 0061 plus a combining diacritic grave 0300 to give *à*, Unicode also contains a precomposed *a*-grave *à* 00E0. I assume this is to help provide backwards compatibility with older encoding systems. Both Macintosh and Windows implementations of Unicode senses when the user enters a combination of base letter and diacritic or diacritics for which a precomposed character exists, and then displays the glyph associated with that precomposed character. This may be a solution to some problems, such as display of stacked diacritics at a time before the operating system (properly speaking, uniscribe) was able to handle stacking, but it can create another problem if the precomposed character glyphs do not have the elements in the desired position. For instance most Americanists would place the ogonek under the belly of the letter *a*, while IPA and Polish prefer the subscript attached to the stem. Most fonts, including the ones on which the YNLIC fonts are based, will follow the IPA/Polish practice. This variation in diacritic placement between precomposed and composed representations may sound minor but in a text with a lot of nasal vowels and superscript diacritics it is readily noticeable. The ogonek will appear in different positions under the *a*, depending on whether or not the precomposition logic can associate the two parts with the precomposed 0105 aogonek.

In the absence of any means within a font to turn off the automatic substitution of precomposed characters, there are perhaps three ways to circumvent the problem. One is to track down all the offending precomposed characters and eliminate them from the font. They will no longer be necessary as appropriate Unicode compliant means are available to achieve suitable display and backing store results. Eliminating all precomposed characters from a Latin Unicode font might substantially decrease its size. But some implementations may substitute a precomposed character from another font if one is not available. A second solution is to fix the glyphs associated with the precomposed characters so that they conform to desired appearance. This is the most labour intensive option but may ultimately be the only workable solution for all implementations. The third way, and the one used here, is to remove the Unicode values from these characters so that they become mere glyphs in the font. This method involves the least modification to the font, and the glyphs are still available if needed at a later date.

### Precomposed characters in the YNLIC fonts with Unicode values removed:

Vowels plus ogonek (no O- or oogonek in Unicode):	E, e, O, o with macron and grave or acute:	T, h, t with macron below
0104 Aogonek	1E14 Emacrongrave	1E6E Tmacronbelow
0105 aogonek	1E15 emacrongrave	1E96 hmacronbelow
0118 Eogonek	1E16 Emacronacute	1E6F tmacronbelow
0119 eogonek	1E17 emacronacute	
012E Iogonek	1E50 Omacrongrave	
012F iogonek	1E51 omacrongrave	
0172 Uogonek	1E52 Omacronacute	
0173 uogonek	1E53 omacronacute	