Marking Tone

1 February 2007
Lorna A. Priest, SIL International, Non-Roman Script Initiative (NRSI)

Abstract:
This paper is intended to explain why pitch contours, which are used in linguistic analysis and documentation of tone languages, have been encoded in the SIL Corporate PUA and in our SIL Unicode Roman fonts. Section 1 covers the existing tone marking systems in Unicode. Section 2 provides the rationale for adding Pitch Contour characters to the PUA, and a detailed discussion of how such characters have been encoded and implemented.

1 Existing Systems in Unicode

Unicode already supports many different systems for marking tone. These are discussed in the following subsections.

1.1 Superscript/Subscript numbers

Superscript numbers indicate a numbered toneme (e.g. 1 = first tone). In Chinese linguistics ^{1} represents low tone and ^{5} represents high tone and Americanists represent ^{1} as high tone and ^{5} as low tone. In African linguistics both orders are found.

Appended numbers give tonal contours directly (e.g. 35 = high rising)

USV: U+00B9, U+00B2, U+00B3, U+2074, U+2075
U+2081, U+2082, U+2083, U+2084, U+2085

Characters: ^{1}, ^{2}, ^{3}, ^{4}, ^{5}, $_{1}$, $_{2}$, $_{3}$, $_{4}$, $_{5}$

Sample Usage: ^{2}ah^{3} há^{2}ah^{3} chia̲á̲h^{1}ah^{3} jnia^{2},

Sample Usage: a$_{1}$jaun$_{2}$ ca$_{2}$rë$_{3}$chán$_{3}$ guein$_{2}$ ne$_{54}$

Issue: cannot be used in identifiers

1.2 Tone Diacritics and Contour tone marks

Diacritics often mark tone in orthographies as well as in linguistic writings. However, this method does not permit one to represent tone phonetically until after one has done phonological analysis.

USV: U+0300, U+0301, U+0302, U+030C, U+0304, U+030D,
U+1DC7, U+1DC5, U+1DC4, U+1DC6, U+1DC8, U+1DC9

Characters: ◌̀, ◌́, ◌̂, ◌̌, ◌̄, ◌̍, ◌᷇, ◌᷅, ◌᷄, ◌᷆, ◌᷈, ◌᷉ (low, high, falling, rising, mid, vert. mid, high mid, low mid, mid-high, mid-low, low-high-low, high-low-high)

Sample Usage: ba᷄t, na᷇pfo tu᷈na, na᷅ mi᷆, bàlónga᷉káé, bo᷈mo᷉támbá

1.3 Modifier letters

Modifier letters are sometimes used to mark tone in orthographies. These are too numerous to discuss all the possibilities. However, two examples are shown.

1.3.1 IPA Downstep/Upstep

IPA includes two symbols to indicate tonal downstep and tonal upstep. The IPA symbols for downstep and upstep are raised, half-height arrows. (IPA also has full-height arrows as distinct symbols, used to represent ingressive versus egressive airflow in disordered speech.)

USV: U+F19D, U+F19C (Proposed codepoints are U+A71C, U+A71B.)

Characters: ꜜ, ꜛ (downstep, upstep)

Sample Usage: yáꜛká, áꜜŋwú̘

1.3.2 Africanist Downstep/Upstep

Africanist linguists have traditionally had their own preferred conventions for indicating downstep and upstep, which are different from the IPA-recommended symbols. Tonal downstep was indicated by using a superscript exclamation mark. For upstep, an inverted exclamation mark was used; in some publications this is superscripted, while in others it was subscripted.

Functionally, the down and up-arrows of the IPA are equivalent to the exclamation and inverted exclamation marks, respectively.

USV: U+F19E, U+F19F, (Proposed codepoints are U+A71D, U+A71E and U+A71F (for subscripted upstep).)

Characters: ꜝ, ꜞ, (downstep, upstep)

Sample Usage: HꜝH, LꜞH, LꜟH, éꜝbéy ꜝméꜝmwét, baatį́ꜞlyá ꜞkį́ꜞndyé

1.3.3 Marking tone in Lahu and Akha

These are orthographic characters which are used for the Lahu and Akha languages of Southeast Asia.

USV: U+02C9, U+02C7, U+02EC, U+02CD, U+02C6, U+F1E7 (proposed codepoint is U+A788)

Characters: ˉ, ˇ, ˬ, ˍ, ˆ, ꞈ

Sample Usage: Ngaˬ -ahˇ hawˬ maˬ mehꞈ nya si…

1.3.4 Marking tone in Chinantec

These are orthographic characters which are used for the Ozumacín Chinantec language of Mexico.

USV: U+02C9, U+02CB, U+02C8, U+02CA, U+A717, U+A718, U+A719, U+A71A

Characters: ˉ, ˋ, ˈ, ˊ, ꜗ, ꜘ, ꜙ, ꜚ

Sample Usage: Jnäꜘ Paaˊ naˉhña̱a̱nˊ la̱a̱nˈ apóstol kya̱a̱ꜗ Jesucristo läꜙ hyohˉ dsëꜗ Dio. Ko̱ˉjø̱hꜘ kya̱a̱hˊ Sóstene ø̱ø̱hꜗ jneˊ,

1.4 Chao tone letters

1.4.1 Right-stem Tone Letters

Each tone letter refers to one of five distinguishable tone levels. To represent contour tones, the tone letters are used in combinations. [TUS 4.0. p 185] This method does not have enough levels to mark the number of levels many African systems need.

USV: U+02E5, U+02E6, U+02E7, U+02E8, U+02E9

Characters: ˥, ˦, ˧, ˨, ˩

Features: with and without staves (this already exists in Doulos SIL)
tone numbers (this already exists in Doulos SIL)
ligated and not-ligated

Sample Usage: ˥ + ˩ = ˥˩

1.4.2 Left-stem Tone Letters (tone sandhi)

In Chinese linguistics, utterances in which tone sandhi occurs are sometimes transcribed using paired tone letters: one right-stemmed tone letter on the left, indicating the underlying tone, and a left-stemmed tone letter on the right, indicating the surface "sandhi" tone. [n2646_CommentsOnN2626.pdf p.9]

USV: U+A712, U+A713, U+A714, U+A715, U+A716

Characters: ꜰ, ꜱ, Ꜳ, ꜳ, Ꜵ

Features: with and without staves
ligated and not-ligated

Sample Usage: ꜵ + Ꜳ + Ꜳ = ꜷ

1.4.3 Dotted tone letters

Dot tone letters are used in Chinese linguistics to indicate tones in certain weakly-stressed syllables having a less-distinct quality—there is little or no pitch variation, and the duration is short. These are often referred to in Chinese linguistics as "neutral tones". [n2646_CommentsOnN2626.pdf p.12] Left-stemmed and dotted tone letters are used contrastively.

USV: U+A708, U+A709, U+A70A, U+A70B, U+A70C, U+A70D, U+A70E, U+A70F, U+A710, U+A711

Characters: ꜈, ꜉, ꜊, ꜋, ꜌, ꜍, ꜎, ꜏, ꜐, ꜑

Features: with and without stems

1.5 Corner tone marks for Chinese

These tone symbols are used by Chinese linguists. Corner tone marks are a distinct transcription tradition from stemmed tone letters.

USV: U+A700, U+A701, U+A702, U+A703, U+A704, U+A705, U+A706, U+A707

Characters: ꜀, ꜁, ꜂, ꜃, ꜄, ꜅, ꜆, ꜇

1.6 Other

Scripts such as Ethiopic, Lao, Thai, Tai Le, New Tai Lue and Nko, all have systems for marking tone. There are a few other characters in the Unicode Standard which have the word tone in their name or their decomposition.

1.6.1 CJK diacritic tone marks

USV: U+302A, U+302B, U+302C, U+302D, U+302E, U+302F

Characters: ◌〪, ◌〫, ◌〬, ◌〭, ◌〮, ◌〯

1.6.2 Departing tone marks

USV: U+02EA, U+02EB

Characters: ˪, ˫

1.6.3 Ethiopic tonal marks

USV: U+1390..U+1399

Characters: ፐ, ፑ, ፒ, ፓ, ፔ, ፕ, ፖ, ፗ, ፘ, ፙ

Sample Usage: የሐፅበ፡በወደ

Application Implementation: "As a calligraphic system of annotation up to this point, there is no known precedence for the rendering of the symbols in typesetting software. It is anticipated however that software capable of handling Kanbun marks, Ruby and similar systems of interlinear annotation, will be readily adaptable to the requirements of Ethiopic tonal notation." [n2747a.pdf] Software will have to know that annotations are to be shrunken down to the 1/4 scale and will shrink everything. This should also make it easier to do layout in 3 rows above a line of text vs the combining mark approach.

1.6.4 Lao

USV: U+0EC8..U+0ECB

Characters: ◌ˈ, ◌ˇ, ◌ˊ, ◌ˋ

1.6.5 New Tai Lue

USV: U+19C8..U+19C9

Characters: ◌᧘, ◌᧙

1.6.6 Nko

USV: U+07EB..U+07F1, U+07F4..U+07F5

Characters: ◌ߋ, ◌ߌ, ◌ߍ, ◌ߎ, ◌ߏ, ◌ߐ, ◌ߑ, ◌ߒ, ◌ߓ

1.6.7 Spacing clones of diacritics

USV: U+02C7, U+02C9, U+02CA, U+02CB, U+02D9

Characters: ◌ˇ, ◌ˉ, ◌ˊ, ◌ˋ, ◌˙

1.6.8 Tai Le

USV: U+1970..U+1974

Characters: ◌ᥰ, ◌ᥱ, ◌ᥲ, ◌ᥳ, ◌ᥴ

1.6.9 Thai

USV: U+0E48..U+0E4B

Characters: ◌่, ◌้, ◌๊, ◌๋

1.6.10 Uralicist tone markers

USV: U+02F9, U+02FA, U+02FB, U+02FC, U+A720, U+A721

Characters: ◌˹, ◌˺, ◌˻, ◌˼, ◌꜠, ◌꜡

1.6.11 Vietnamese tone marks (deprecated)

USV: U+0340, U+0341

Characters: ◌̀, ◌́

1.6.12 Zhuang orthographic tones

USV: U+0184, U+0185, U+01A7, U+01A8, U+01BC, U+01BD

Characters: Ƅ, ƅ, Ƨ, ƨ, Ƽ, ƽ

2 9-level pitch contours

Existing tone systems in Unicode distinguish up to 5 levels of tone. However, there is a 9-level system, used mainly by Africanists. The nine level system includes contour pitch marks and is used for phonetic transcriptions of tone languages. For this reason these characters are referred to as pitch rather than tone contours. This system is widely used by Africanists, not only for transcribing raw phonetic data prior to tone analysis, but for archiving and producing linguistic documents. Smart-font tables within a Graphite or OpenType font would substitute glyphs for angled bars indicating pitch contours based on what sequence of pitch-level characters occur together. Initially we have encoded the nine pitches in our Private Use Area (PUA). At some point a Unicode proposal would need to be made. Even after characters are approved for addition to Unicode it can take up to two years to officially get into the Unicode standard and then a year or two later before Microsoft would implement it in Uniscribe. Thus, implementations in the near future needs to use PUA characters and be implemented in Graphite fonts for use in software that supports Graphite rendering.

2.1 Encoding and Properties

The pitch contours are modifier letters; they are not combining marks. In the event that these characters do get accepted into Unicode, users should not expect that the USVs will bear any relation to the pitch numbers (as they do now).

USV: U+F1F1, U+F1F2, U+F1F3, U+F1F4, U+F1F5, U+F1F6, U+F1F7, U+F1F8, U+F1F9

Characters: _, _, –, –, –, –, ¯, ¯, ¯

Sample Usage:
/ _ – – –
[took go rɪ gba tuk] ‘colleague is looking at bushbaby’

Sample Usage:
[– _ \ ﹀/\ /\ ]
mɑ mɑmɑ t^{h}ədaⁱəzsat^{h}ədaⁱ jɛs

Sample Usage:
– _\ _/‾\ /\
mɑ mɑmɑ t^{h}ədaⁱəzsat^{h}ədaⁱ jɛs

Sample Usage (inline): L lə̀tɔ̀ŋ [_ ˎ] *feather*

Properties:

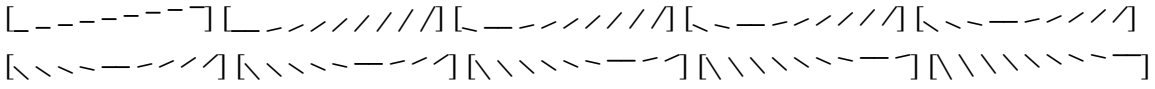
```
F1F1;MODIFIER LETTER PITCH ONE;Sk;0;ON;;;;;N;;;;;
F1F2;MODIFIER LETTER PITCH TWO;Sk;0;ON;;;;;N;;;;;
F1F3;MODIFIER LETTER PITCH THREE;Sk;0;ON;;;;;N;;;;;
F1F4;MODIFIER LETTER PITCH FOUR;Sk;0;ON;;;;;N;;;;;
F1F5;MODIFIER LETTER PITCH FIVE;Sk;0;ON;;;;;N;;;;;
F1F6;MODIFIER LETTER PITCH SIX;Sk;0;ON;;;;;N;;;;;
F1F7;MODIFIER LETTER PITCH SEVEN;Sk;0;ON;;;;;N;;;;;
F1F8;MODIFIER LETTER PITCH EIGHT;Sk;0;ON;;;;;N;;;;;
F1F9;MODIFIER LETTER PITCH NINE;Sk;0;ON;;;;;N;;;;;
```

2.2 Implementation

2.2.1 Position

The lowest pitch is below the baseline; it is at the level of the lowest arm of a square bracket.

The highest pitch is slightly above cap height; it is at the level of the top arm of a square bracket.



2.2.2 Smart fonts

The font should contain Graphite rules for converting a sequence of the nine-pitches to the contours. For example U+F1F1 U+F1F9 U+F1F4 U+F1F7 would be rendered as:  . Following Chao notation, two of the same pitch bars, such as <U+F1F5 U+F1F5>, render as a double-length bar.

	1	2	3	4	5	6	7	8	9
1									
Encoding	U+F1F1 U+F1F1	U+F1F1 U+F1F2	U+F1F1 U+F1F3	U+F1F1 U+F1F4	U+F1F1 U+F1F5	U+F1F1 U+F1F6	U+F1F1 U+F1F7	U+F1F1 U+F1F8	U+F1F1 U+F1F9
2									
Encoding	U+F1F2 U+F1F1	U+F1F2 U+F1F2	U+F1F2 U+F1F3	U+F1F2 U+F1F4	U+F1F2 U+F1F5	U+F1F2 U+F1F6	U+F1F2 U+F1F7	U+F1F2 U+F1F8	U+F1F2 U+F1F9
3									
Encoding	U+F1F3 U+F1F1	U+F1F3 U+F1F2	U+F1F3 U+F1F3	U+F1F3 U+F1F4	U+F1F3 U+F1F5	U+F1F3 U+F1F6	U+F1F3 U+F1F7	U+F1F3 U+F1F8	U+F1F3 U+F1F9
4									
Encoding	U+F1F4 U+F1F1	U+F1F4 U+F1F2	U+F1F4 U+F1F3	U+F1F4 U+F1F4	U+F1F4 U+F1F5	U+F1F4 U+F1F6	U+F1F4 U+F1F7	U+F1F4 U+F1F8	U+F1F4 U+F1F9
5									
Encoding	U+F1F5 U+F1F1	U+F1F5 U+F1F2	U+F1F5 U+F1F3	U+F1F5 U+F1F4	U+F1F5 U+F1F5	U+F1F5 U+F1F6	U+F1F5 U+F1F7	U+F1F5 U+F1F8	U+F1F5 U+F1F9
6									
Encoding	U+F1F6 U+F1F1	U+F1F6 U+F1F2	U+F1F6 U+F1F3	U+F1F6 U+F1F4	U+F1F6 U+F1F5	U+F1F6 U+F1F6	U+F1F6 U+F1F7	U+F1F6 U+F1F8	U+F1F6 U+F1F9
7									
Encoding	U+F1F7 U+F1F1	U+F1F7 U+F1F2	U+F1F7 U+F1F3	U+F1F7 U+F1F4	U+F1F7 U+F1F5	U+F1F7 U+F1F6	U+F1F7 U+F1F7	U+F1F7 U+F1F8	U+F1F7 U+F1F9
8									
Encoding	U+F1F8 U+F1F1	U+F1F8 U+F1F2	U+F1F8 U+F1F3	U+F1F8 U+F1F4	U+F1F8 U+F1F5	U+F1F8 U+F1F6	U+F1F8 U+F1F7	U+F1F8 U+F1F8	U+F1F8 U+F1F9
9									
Encoding	U+F1F9 U+F1F1	U+F1F9 U+F1F2	U+F1F9 U+F1F3	U+F1F9 U+F1F4	U+F1F9 U+F1F5	U+F1F9 U+F1F6	U+F1F9 U+F1F7	U+F1F9 U+F1F8	U+F1F9 U+F1F9

2.2.3 Preventing ligation of adjacent characters

A space can be used to prevent ligation of adjacent characters. If no space is wanted between characters U+200C ZWNJ could be used. A problem will happen if a program does not properly interprets ZWNJ and users try other ways to get the appearance they want. An opening or closing bracket should never be put on a separate line, for example. If a vertical bar U+0070 or double bar U+2016 are used to separate intonation groups, line breaking should work properly (probably to break after the bar). In this case U+2060 WORD JOINER may be appropriate and necessary before the bar.

2.2.4 Font features

The features we have implemented in SIL's Unicode Roman fonts are

→without (default) and with tramlines (if brackets are desired, the user would type them in). A space (U+0020) is the only other character we have implemented tramlines for.

→ligated (default) and non-ligated

Converting data between tone-bar or pitch-contour notation and superscript-numeral notation should be considered as a change to the underlying text, not a glyph alternate for some "abstract" pitch character. It would be similar to converting between IPA and "Americanist" phonetic notation, or between Latin and Cyrillic orthographies for a Central Asian language, etc. Thus, we believe that TECkit mapping files should be created (these have not yet been created) to convert between the different systems and font features should not be used. This is discussed in Section 3.

2.2.5 Font issues

In order to allow for the largest size possible for the pitch contours we have placed them as far down as is reasonable (at the bottom arm of the square bracket) and as high as possible (at the top arm of the square bracket). This implementation should not interfere with other uses for the font. If larger pitch contours are desired, the point size can be changed in a document.

The tones and pitch contours are all the same width. However, because they are in proportional fonts there will always be the problem of horizontal alignment of the pitch contours with segmental symbols on a line above or below. Software solutions to this alignment issues could include treating them as interlinear text.

Application Implementation: The application should take care of any visual adjustments such as bracket height and size, interlinearization, enlarging the pitches. See also suggestions under Ethiopic Tonal Marks concerning Ruby notation.

2.2.6 Keyboarding

Only the nine levels of pitch need keystrokes. The "contours" will render automatically in Graphite-enabled applications because of the smart-font tables. We have implemented keyboarding for these pitches in the "IPA Unicode 1.1 Keyman Keyboard" (http://scripts.sil.org/UniIPAKeyboard)

2.3 Publications which use the bar notation

Anderson, Stephen R. 1978. *Tone features.* In Victoria A. Fromkin (ed.). Tone: A Linguistic Survey. New York: Academic Press, pp. 133-175.

Clark, Mary. 1993. *Representation of downstep in Dschang Bamileke.* In Harry Van der Hulst and Keith Snider (eds.). The Phonology of Tone: The Representation of Tonal Register. Berlin: Mouton de Gruyter, pp. 31, 39.

Hyman, Larry M. 1979. *A reanalysis of tonal downstep.* Journal of African Languages and Linguistics 1.1:9-29.

Laughren, Mary. 1984. *Tone in Zulu nouns.* In Autosegmental Studies in Bantu Tone edited by G.N. Clements & John Goldsmith. Dordrecht: Foris Publications, pp. 223-225.

Pulleyblank, Douglas. 1986. *Tone in Lexical Phonology.* Dordrecht: D. Reidel Publishing Co., pp. 27, 28, 30, 31, to name a few places.

Snider, Keith. 1990. *Tonal upstep in Krachi: evidence for a register tier.* Language 66.3: 453-474.

Snider, Keith. 1998. *Phonetic realisation of downstep in Bimoba.* Phonology 15.1: 77-101.

Yip, Moira. 2002. *Tone.* Cambridge: Cambridge University Press. Various places throughout when needed, e.g., pp. 149, 150, 266.

3 Data conversion

We should be able to create TECkit mappings that would be able to transform data between different types of transcription methods, including tone bars. This has not yet been done.

→superscript/subscript numbers ↔ tone bars

→superscript/subscript numbers ↔ pitch contours