To: Unicode Technical Committee **L2/04-214**
From: Deborah Anderson
Date: 7 June 2004
RE: Request to consider superscripts and subscripts in linguistics

1 Overview and Rationale

This is a request for the UTC to discuss the subject of superscripts and subscripts, and in particular to suggest when superscripts/subscripts characters should be used, as opposed to style. These are issues relevant to the proposed subscripts in L2/04-191 (N2788).

For those in Indo-European, the style capability has made it easy for users to create subscripts and superscripts at will, without having to think about the difference between plain and rich text or whether a particular character is in Unicode or not. Many users I queried rely regularly on the style feature for sub/superscripts, unaware of its impact on search features and that during the electronic exchange of documents such information could be lost.

The examples below are drawn from Indo-European and more general linguistics. Feedback from the UTC is needed in order for me to better counsel linguistic projects on how to handle superscripts/subscripts and when new characters ought to be proposed. It also will provide guidance on developing a Unicode font for Indo-European linguists, currently underway, which may help establish a more consistent usage of encoded characters (vs. style).

2 Repertoire of Current Missing Latin Letters as Subscripts and Superscripts

Nearly all small Latin superscript letters have been accepted (or are proposed), the only exception being lowercase q. Eight capital superscript letters are missing. Four small Latin letters have been encoded as subscripts, four are proposed, and no capital Latin letters subscripts have been encoded. (In the following * indicates that the character has been proposed in L2/04-132.)

Missing superscripts:
Small Latin letters: c*, f*, q, z*
Capital letters: C, F, Q, S, V, X, Y, Z

[Also included are the following lower case Greek letters: β, γ, δ, φ, χ.
Modifier small capital letter I, L, N, and U are all proposed in L2/04-132.]

Missing subscripts:
Small Latin letters: a*, b, c, d, e*, f, g, h, j, k, l, m, n, o*, p, q, s, t, w, x*, y, z
All capital Latin letters are missing
(* proposed in L2/04-191)
[Also included are: β, γ, ρ, φ, χ]

3 Subscripts and Superscripts used in IE

Superscripts and subscripts can occur in Proto-Indo-European reconstructions, including phonetic and phonemic transcription of reconstructed forms. Some of the most commonly occurring sub/superscripts include:

spacing modifier letters ^{h} and ^{w} (or, from Phonetic Extensions, ^{u} with combining inverted breve below)

subscripts 1, 2, 3, 4, a , e , o, x, and ə (the latter five are proposed in L2/04-191)
(Note that the notation h_1 is the same as h_e, $h_2 = h_a$, and $h_3 = h_o$)

Other subscripts and superscripts used for transliteration and transcription include, amongst others:

subscripts: ä, i, u for Tocharian

superscript: lowercase d, f, mi and hi (h usually with combining breve below), etc. in Hittite
most capitals (some with diacritics, occasionally a superscript *with* a subscript, all used for Sumerograms in Hittite see 6d, below)

Still, the encoded sub/superscript characters do not cover all the sub/superscripted characters that appear in Indo-European publications. I believe the problem has not surfaced because the style feature has been widely used in order to make any character sub/superscripts needed.

3 Styled Text?

In general, sub/superscripts that appear in reconstructed forms are set in italicized type, but not always:

- Those subscripts and superscripts that occur in phonemic and phonetic transcriptions of Proto-Indo-European (or, below, pre-Indo-European) are usually set in plain text

It developed in patterns when the subject directly preceded a finite verb. Conditions were then right for lengthened grade; for with the loss of accent on a following verb, its stem vowel dropped out and the preceding vowel was lengthened; for example, pre-Indo-European /pₑtér/ + /éyty/ became /pₑté·r yty/.

Figure 1. W. P. Lehmann, "On Earlier Stages of the Indo-European Nominal Inflection," Language, 34.2. (Apr. - Jun., 1958), p 198.

- Some scholars, such as Don Ringe at U of Penn, regularly does not italicize reconstructed forms. The second example below is from an article by V. Shevoroshkin, who also does not italicize (except for the "w", for some reason).

1. *x́, *x, *x^{w} are the "laryngeals", usually written *h_1, *h_2, *h_3 respectively; their position in the phonemic chart is frankly speculative (cf. Beekes 1972b:45).

Figure 2. Don Ringe Jr,. *On the Chronology of Sound Changes in Tocharian*, vol. 1: *From Proto-Indo-European to Proto-Tocharian*, New Haven, 1996

For H. Eichner, N. Oettinger and other young scholars belonging to Eichner's school the idea of vowel-coloring laryngeals is still alive: they still reconstruct "*h_1"(= *h́), "*h_3" (= *h^{w}), though there is no reason for it. It is

Figure 3. V. Shevoroshkin, "On Laryngeals", in Die Laryngaltheorie, ed. A. Bammesberger, Heidelberg 1988, p. 528, with un-italicized subscript numbers (but "w" italicized, for some reason).

4 Fonts

Many current fonts used by many Indo-Europeanists are not Unicode compliant (such as TimesIndogermania and IEPalatino). The fonts often include the glyphs for numerical subscripts and occasionally other sub/superscripts, but a brief survey of users indicates that the "style" function is commonly used, since this allows any character to become a sub/superscript.

5 Searching and Typing

It is possible to search for superscript styled letters and mixed combinations of plain and styled text in MS Word. (I do not know whether the style distinctions are regularly picked up by search processes or how well they work when exchanging documents between word processing programs.)

Typing superscripted style is more difficult than typing a plain text stream. Yet, switching constantly between style and typing letters (as independent characters) can be cumbersome and confusing.

6 General Categories of Sub/Superscripting Use

Sub/superscripts are used in Indo-European and general linguistics materials in a variety of ways. Examples are provided below which may help guide the creation of guidelines.

a. Describing an entity or defining a law or rule

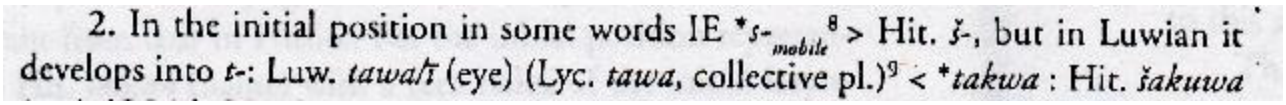

2. In the initial position in some words IE *s_{mobile}[8] > Hit. *š-*, but in Luwian it develops into *t-*: Luw. *tawa/ī* (eye) (Lyc. *tawa*, collective pl.)[9] < **takwa* : Hit. *šakuwa*

Figure 4. Vyacheslav Ivanov, "South Anatolian and Northern Anatolian," in *Greater Anatolia and the Indo-Hittite Language Family,* ed. R. Drews, *JIES* Monograph 38, Washington DC, 2001, p. 133

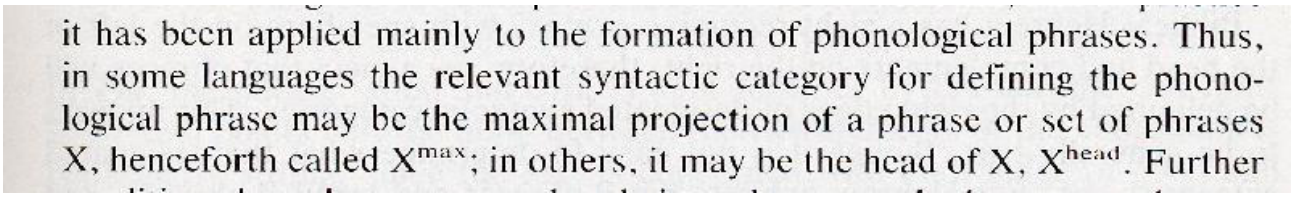

it has been applied mainly to the formation of phonological phrases. Thus, in some languages the relevant syntactic category for defining the phonological phrase may be the maximal projection of a phrase or set of phrases X, henceforth called X^{max}; in others, it may be the head of X, X^{head}. Further

Figure 5. Bezalel Elan Dresher, "Tiberian Hebrew System of Accents" in *Language*, 7.1 (March 1994), p. 17

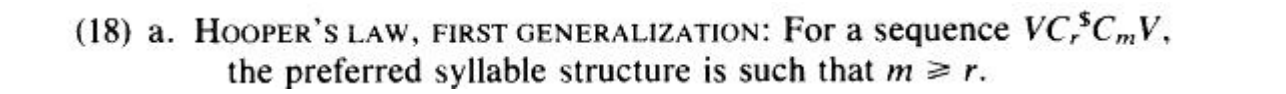

(18) a. HOOPER'S LAW, FIRST GENERALIZATION: For a sequence $VC_r{}^{\$}C_mV$, the preferred syllable structure is such that $m \geq r$.

Figure 6. Theo Vennemann and Robert Murray, "Sound Change and Syllable Structure in Germanic Phonology," Language p. 519, Language, 59.3, 1983

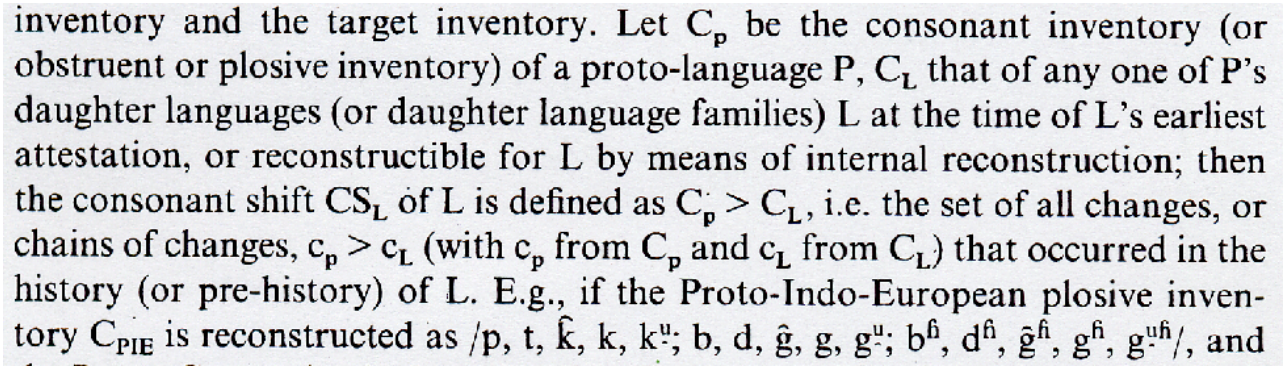
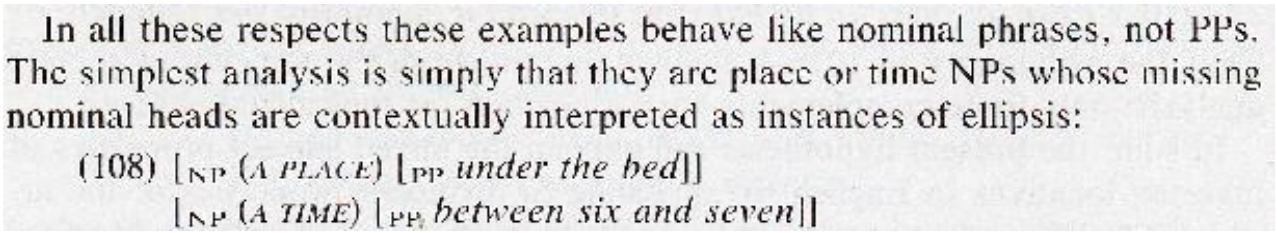

inventory and the target inventory. Let C_p be the consonant inventory (or obstruent or plosive inventory) of a proto-language P, C_L that of any one of P's daughter languages (or daughter language families) L at the time of L's earliest attestation, or reconstructible for L by means of internal reconstruction; then the consonant shift CS_L of L is defined as $C_p > C_L$, i.e. the set of all changes, or chains of changes, $c_p > c_L$ (with c_p from C_p and c_L from C_L) that occurred in the history (or pre-history) of L. E.g., if the Proto-Indo-European plosive inventory C_{PIE} is reconstructed as /p, t, k̂, k, k^u; b, d, ĝ, g, g^u; $b^ɦ$, $d^ɦ$, $ĝ^ɦ$, $g^ɦ$, $g^{uɦ}$/, and

Figure 7. Theo Vennemann, in *The New Sound of Indo-European*, ed. Theo Vennemann, Trends in Linguistics. Studies and Monographs 41, Berlin & New York: Mouton de Gruyter, 1989

b. Use of subscripted "NP," "VP," etc., in general linguistics texts

"NP" for "Noun Phrase" is used with layout features more reminiscent of math than plain text. In the journal *Language* these are set off from the main text. I would recommend the following cases be handled with style.

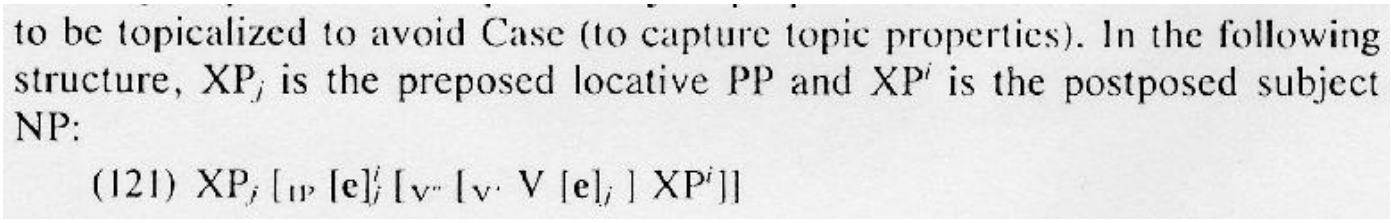
In the following example, "NP" (noun phrase) and "PP" (prepositional phrase) are subscripted



In all these respects these examples behave like nominal phrases, not PPs. The simplest analysis is simply that they are place or time NPs whose missing nominal heads are contextually interpreted as instances of ellipsis:

(108) [$_{NP}$ (*A PLACE*) [$_{PP}$ *under the bed*]]
[$_{NP}$ (*A TIME*) [$_{PP}$ *between six and seven*]]

Figure 8. Joan Bresnan, "Locative Inversion and the Architecture of Universal Grammar" in *Language*, 7.1 (March 1994), p. 110



to be topicalized to avoid Case (to capture topic properties). In the following structure, XP_j is the preposed locative PP and XP^i is the postposed subject NP:

(121) XP_j [$_{IP}$ [e]^i_j [$_{V''}$ [$_{V'}$ V [e]$_j$] XP^i]]

Figure 9. Joan Bresnan, "Locative Inversion and the Architecture of Universal Grammar" in *Language*, 7.1 (March 1994), p. 120

c. Used in phonetic/phonemic transliteration and in reconstruction

subscript c (IH indicates "Indo-Hittite"):

IH '*ъdontes* or /'dontes/ = ['$_c$dontes, '$_v$dontes]

Figure 10. Paul Brosman, Jr. "Proto-Indo-Hittite ъ and the Allophones of Laryngeals" *Language* 33.1, Jan-Mar 1957, p. 3

subscript front jer (ь):

the previous ones, for the second phoneme is not a consonant but a sonant. In the same sentence collocation as that which produced a vowel in the other forms, we would have, according to the formula for two sonants after a consonant and before a vowel, either H_v*wos-* or H_c*uwos-*, depending upon what preceded the final consonant. In the case of the second of these doublets, that in which the

Figure 11. Paul Brosman, Jr. "Proto-Indo-Hittite jer and the Allophones of Laryngeals" *Language* 33.1, Jan-Mar 1957, p. 4

d. Cuneiform Transliteration

Another area where superscripting has been used extensively is in cuneiform transliteration. Capital letters are used to designate Sumerian determinatives and plural markers. (Akkadian

capital letters, used for phonetic complements, are also capitalized, but are italicized.)[1] Because eight capital superscripts letters are missing, this would limit the possibilities of what could be spelled, if determinatives and phonetic complements/plural markers were not handled with style.

Examples:
Sumerian Determinative:

^{DUG}harharan '*harharan*'-vessel (^{DUG} is a Sumerogram for 'vessel', *harharan* is a Hittite word for a particular kind of vessel)

^{GIŠ}BANŠUR '[wooden] table' (^{GIŠ} is a Sumerogram for 'wood', it occurs before wooden objects, here a table)

$^{NA}{}_4$ *peruna*- 'rock' ($^{NA}{}_4$ is a Sumerogram indicating an element made of minerals, *peruna*- is the Hittite word; properly, "4" should be raised higher as it is a subscript to the superscript "NA")

Lowercase superscripts (d, f, and m) are also used for determinatives:

^{d}IM 'Storm God' (^{d} is a determinative, short for DINGIR, indicating a deity, here the Storm God; the captital letter ^{D} is also possible)

Sumerian plural marker:

DINGIR^{MEŠ} 'gods' ("DINGIR" is Sumerogram for 'god' + Sumerian plural marker, ^{MEŠ})

Craig Melchert reports there is no consensus on superscripting for plural markers after Sumerograms (or Akkadian phonetic complements); some publications and scholars do superscript these, others do not. The *Chicago Hittite Dictionary*, for example, does not superscript plural markers after Sumerograms: DINGIR.MEŠ.

This inconsistency suggests that the need to finish encoding all the superscript letters for cuneiform transliteration is not warranted, at least for plural markers (and Akkadian complements). It may be that the superscripted determinatives are less problematic and there is more consistency. According to Craig Melchert, an expert on Hittite and other Anatolian scripts at the University of North Carolina, Chapel Hill, the use of superscripting with determinatives is done to indicate something that is not phonetically real, in other words, the represented elements do not stand for anything in the spoken chain. In some cases, however, it is not always clear whether the elements are spoken or not, so there can uncertainty on

[1] An example of a Akkadian phonetic complement would be: DINGIR^{LIM} 'god' ("DINGIR"is a Sumerogram for 'god' with the Akkadian phonetic complement ^{LIM}). Note that the *Chicago Hittite Dictionary* also does not superscript Akkadian complements: it uses DINGIR-*LIM* .

whether to superscript a particular Sumerogram or not. Regarding the topic of superscripting, Steve Tinney writes: “I do know that I have come to the conclusion that life without superscript characters to handle determinatives at least is very inconvenient.” Further study is needed.

Note: Additional new uses for superscripts have been created in Hittite: Craig Melchert and Harry Hoffner will be using superscripted “hi” and “mi” to indicate the conjugation of a particular Hittite verb in their forthcoming grammar. In this case, the lowercase h, m, and i are already included as superscripts. However, I think this is indicative of the sort of additional use of superscripting and possibly subscripting that will continue to be used by scholars.