**ISO/IEC JTC1/SC2/WG2 N4212**

**L2/11-375**

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

**Doc Type: Working Group Document**
**Title: Proposal to encode combining decimal digits above in the UCS**
**Source: Abteilung für Griechische und Lateinische Philologie der Ludwig-Maximilians-Universität Müchen (Department of Greek and Latin Philology, Ludwig-Maximilians-University of Munich, Germany)**
**Authors: Martin Schrage, Karl Pentzlin**
**Status: Expert Contribution**
**Action: For consideration by JTC1/SC2/WG2 and UTC**
**Date: 2011-10-15**

*This document is based on an excerpt of WG2 N3913 and L2/10-358R (Proposal to encode Metrical Symbols and related characters), as it was decided to split that work and to propose the "related characters" separately by subject.*

## 1. Introduction

In linguistic works, it is sometimes necessary to mark one or more individual letters within a word by decimal numbers placed above them.

Purposes may be:

- to reference the letter (or the sound it denotes, or the syllable it is part of) in subsequent text (a specific example is the numbering of the parts of verses in textbooks).
- to identify such elements of words in subsequent examples where parts occur in another order, due to grammatical or syntactical transformations.
- to denote numeric values of any properties applying to that element.

For the same purposes, this also occurs with metrical symbols within sequences of these.

While in most instances such numbers are below 10 and therefore can be represented by single "combining digits above", larger numbers (requiring 2 decimal digits) occur.
It is addressed below (in section 2.1) how this is handled using such combining digits.

Thus, this document proposed 10 decimal "combining digits above" (0…9).

## 2. Encoding Considerations

The "combining digits above" are proposed to be placed in the next free column (according to PDAM 1.2, WG2 N4107) of the "Combining Diacritical Marks Extended" block, contiguously and in order according to current principles regarding decimal digits. It seems appropriate to start a new column, although this leaves a gap of one code point (which will presumably be filled in the near future anyway).

While they are declared as "decimal digits" by the proposed properties, no compatibility equivalences are given, as this seems not appropriate for combining characters (while such are given for the superscript and subscript digits U+00B2,U+00B3, U+00B9, U+2070, U+2074…U+2079, U+2080…U+2089). Also, this would make text like "${}^{238}_{92}$U" (see below, section 2.2) equivalent to "29328U", which would make no sense.

## 2.1 Side-by-side placing to allow combining decimal numbers ≥ 10

Sequences of "combining digits above" are to be placed side-by-side, rather than be stacked vertically as sequences of combining characters with the combining class otherwise do.

This is compliant with other exceptions listed in "The Unicode Standard, Version 6.0", p.83, section 3.6 "Combination", subheader "Combining Character Sequences", Guideline P3 "Side-by-side application".
It is proposed that the "combining digits above" are listed there also at the time when they are valid Unicode characters.

Thus, combining decimal numbers consisting of more than one number may be displayed using the "combining digits above".

If such is considered necessary, the maximum number of "combining digits above" placed side-by-side may be restricted, even to 2 (as we did not found combining decimal numbers ≥ 100).

## 2.2 Using of "combining digits above" to denote nuclides

Having "combining digits above" encoded, they provide a way to denote nuclides and isotopes in nuclear physics (like $^{238}_{92}\mathrm{U}$) in plain text (independent of the issue whether this possibility alone would justify an encoding).

This is done by placing a single "combining digit above" on the digit which is placed beneath it, for which a "subscript digit" is used. If there is no digit under it, a "no-break space" is used instead. (Thus, the number above is composed from the same digit glyph set, rather than using a "superscript digit" which may not have the same dimensions in a given font. See fig. 2007a-757.)

Thus, $^{238}_{92}\mathrm{U}$ (the common denoting of the most frequent isotope of uranium) is written as follows:

| | |
|---|---|
| U+00A0 | NO-BREAK SPACE |
| (proposed) | COMBINING DIGIT TWO ABOVE |
| U+2089 | SUBSCRIPT NINE |
| (proposed) | COMBINING DIGIT THREE ABOVE |
| U+2082 | SUBSCRIPT TWO |
| (proposed) | COMBINING DIGIT EIGHT ABOVE |
| U+0055 | LATIN CAPITAL LETTER U |

As otherwise combining marks are adjusted to the position and size of the base characters if these are superscript or subscript characters (like when placing a combining diaeresis onto a "modifier letter small a" to get "$^{\ddot{a}}$"), an exception to this rule has to be stated that "combining digits above" placed over a "subscript digit" shall inherit the shape and size from the "subscript digit", rather than to get a smaller size to retain the proportions to the base character, as they are appropriate when being applied to a full-sized character.

It is noted that in such a sequence, the term "238" is not found by a simple substring search. However, such searches can be successful by using somewhat more advanced search algorithms. They are no more complicated than algorithms which have to find base text sequences where diacritical marks spanning over three letters are applied, which are encoded as parts like outlined in the recently accepted proposal WG2 N4078 " Revised Proposal to enable the use of Combining Triple Diacritics in Plain Text".

Also, it is noted that numbers stacked this way within plain text occur also in other areas than physics, even in linguistics (see fig. 2001a-119).

## 3. Proposed Characters

### *Block: Combining Diacritical Marks Extended*

### Combining Digits

| | | |
|---|---|---|
| $\overset{0}{◌}$ | U+1AC0 | COMBINING DIGIT ZERO ABOVE |
| $\overset{1}{◌}$ | U+1AC1 | COMBINING DIGIT ONE ABOVE |
| $\overset{2}{◌}$ | U+1AC2 | COMBINING DIGIT TWO ABOVE |
| $\overset{3}{◌}$ | U+1AC3 | COMBINING DIGIT THREE ABOVE |
| $\overset{4}{◌}$ | U+1AC4 | COMBINING DIGIT FOUR ABOVE |
| $\overset{5}{◌}$ | U+1AC5 | COMBINING DIGIT FIVE ABOVE |
| $\overset{6}{◌}$ | U+1AC6 | COMBINING DIGIT SIX ABOVE |
| $\overset{7}{◌}$ | U+1AC7 | COMBINING DIGIT SEVEN ABOVE |
| $\overset{8}{◌}$ | U+1AC8 | COMBINING DIGIT EIGHT ABOVE |
| $\overset{9}{◌}$ | U+1AC9 | COMBINING DIGIT NINE ABOVE |

### Properties:

```
U+1AC0 COMBINING DIGIT ZERO ABOVE;Mn;230;NSM;;0;0;0;N;;;;;
U+1AC1 COMBINING DIGIT ONE ABOVE;Mn;230;NSM;;1;1;1;N;;;;;
U+1AC2 COMBINING DIGIT TWO ABOVE;Mn;230;NSM;;2;2;2;N;;;;;
U+1AC3 COMBINING DIGIT THREE ABOVE;Mn;230;NSM;;3;3;3;N;;;;;
U+1AC4 COMBINING DIGIT FOUR ABOVE;Mn;230;NSM;;4;4;4;N;;;;;
U+1AC5 COMBINING DIGIT FIVE ABOVE;Mn;230;NSM;;5;5;5;N;;;;;
U+1AC6 COMBINING DIGIT SIX ABOVE;Mn;230;NSM;;6;6;6;N;;;;;
U+1AC7 COMBINING DIGIT SEVEN ABOVE;Mn;230;NSM;;7;7;7;N;;;;;
U+1AC8 COMBINING DIGIT EIGHT ABOVE;Mn;230;NSM;;8;8;8;N;;;;;
U+1AC9 COMBINING DIGIT NINE ABOVE;Mn;230;NSM;;9;9;9;N;;;;;
```

## 4. Acknowledgements

Special thanks to Google for providing scans of numerous 19th century books and publications accessible on the Internet.

## 5. References

[1834a] Munk, Edward: Metrik der Griechen und Römer. – Glogau/Leipzig 1834

[1856a] Anthon, Charles: A System of Latin Prosody and Metre. – New York 1856

[1968a] Korzeniewski, Dieter: Griechische Metrik. – Darmstadt 1968

[1968b] Thummer, Erich: Pindar, die isthmischen Gedichte. – Heidelberg 1968

[1977a] Sovijärvi, Antti, & Reino Peltola, eds. 1977. Suomalais-ugrilainen tarkekirjoitus. (Helsingin Yliopiston Fonetiikan Laitoksen Julkuaisua; 9) – Helsinki: Publicationes Instituti Phonetici Universitatis Helsingiensis. – ISBN 951-45-1019-4

[1993a] Sicking, C. M. J.: Griechische Verslehre. – München 1993, ISBN 3 406 35252 9

[1997a] Nesselrath, Heinz-Günther: Einleitung in die griechische Philologie – Leipzig/Stuttgart 1997, ISBN-13: 978-3519074359

[2001a] Girdenis, Aleksas: Kalbotyros darbai (Studies in linguistics): straipsniai, studijos, esė, recenzijos. T. 3: 1988–2000. – Vilnius, Mokslo ir enciklopedijų leidybos inst., 2001. ISBN 5-420-01480-7

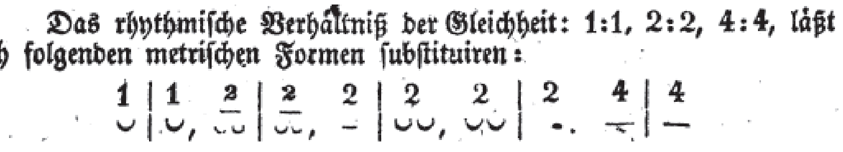
[2007a] Carroll, Bradley W., et al.: An introduction to modern astrophysics. – 2nd ed., San Francisco 2007, ISBN 0-8053-0402-9

## 6. Examples and Figures

The figures are numbered by the referenced work (consisting of the year of edition and the letter, as in the "references" list, followed by a hyphen the page number, and following by a second letter if more than one figure is taken from a page.
E.g.: "Fig. 1834a-7" means "See ref. [1834a], p.7").

For shortness, the combining digits above are referenced simply by their numeric value in the figure legends.

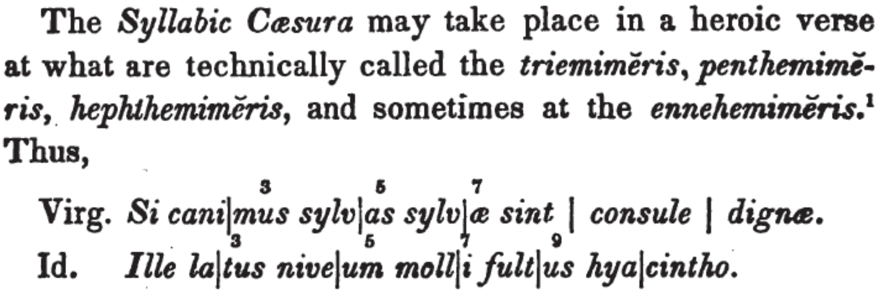
**Fig. 1834a-7:** *Showing specimens for 1, 2, 4, applied to metrical symbols (the pair of breves beneath some of the "2" examples is considered being U+23B6 METRICAL TWO SHORTS JOINED).*



Das rhythmiſche Verhältniß der Gleichheit: 1:1, 2:2, 4:4, läßt ſich folgenden metriſchen Formen ſubſtituiren:

1 | 1 2 | 2 2 | 2 2 | 2 4 | 4
⏑ | ⏑, ⏑⏑ | ⏑⏑, – | ⏑⏑, ⏑⏑ | –. – | —

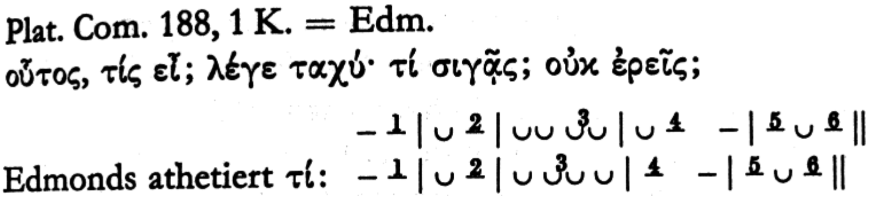
**Fig. 1856a-102:** *Showing specimens for3, 5, 7, 9, applied to Latin letters.*



The *Syllabic Cæsura* may take place in a heroic verse at what are technically called the ***triemimĕris***, ***penthemimĕris***, ***hephthemimĕris***, and sometimes at the ***ennehemimĕris***.[1] Thus,

Virg. *Si cani|mus sylv|as sylv|æ sint | consule | dignæ.* (3, 5, 7)
Id. *Ille la|tus nive|um moll|i fult|us hya|cintho.* (3, 5, 7, 9)

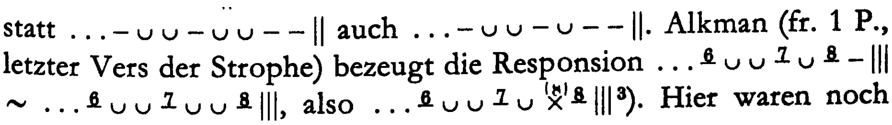
**Fig. 1968a-61:** *Showing specimens for 1...6, applied to metrical symbols.*



Plat. Com. 188, 1 K. = Edm.
οὗτος, τίς εἶ; λέγε ταχύ· τί σιγᾷς; οὐκ ἐρεῖς;

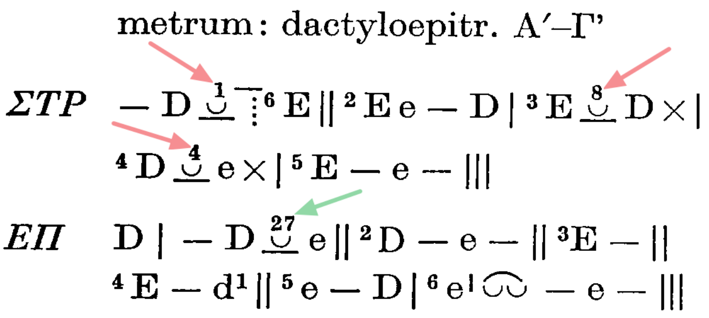
– 1 | ⏑ 2 | ⏑⏑ 3⏑ | ⏑ 4 – | 5 ⏑ 6 ||
Edmonds athetiert τί: – 1 | ⏑ 2 | ⏑ 3⏑ ⏑ | 4 – | 5 ⏑ 6 ||

**Fig. 1968a-74:** *Showing specimens for 1...8, applied to metrical symbols. The 8 in parentheses above the anceps in the last row can be encoded by the sequence COMBIBING DIGIT EIGHT ABOVE + U+1ABB COMBINING PARENTHESES ABOVE (the latter being contained in PDAM1.2, see WG2 N4107).*



statt . . . – ⏑ ⏑ – ⏑̈ ⏑ – – || auch . . . – ⏑ ⏑ – ⏑ – – ||. Alkman (fr. 1 P., letzter Vers der Strophe) bezeugt die Responsion . . . 6 ⏑ ⏑ 7 ⏑ 8 – ||| ∼ . . . 6 ⏑ ⏑ 7 ⏑ ⏑ 8 |||, also . . . 6 ⏑ ⏑ 7 ⏑ (8)× 8 |||3). Hier waren noch

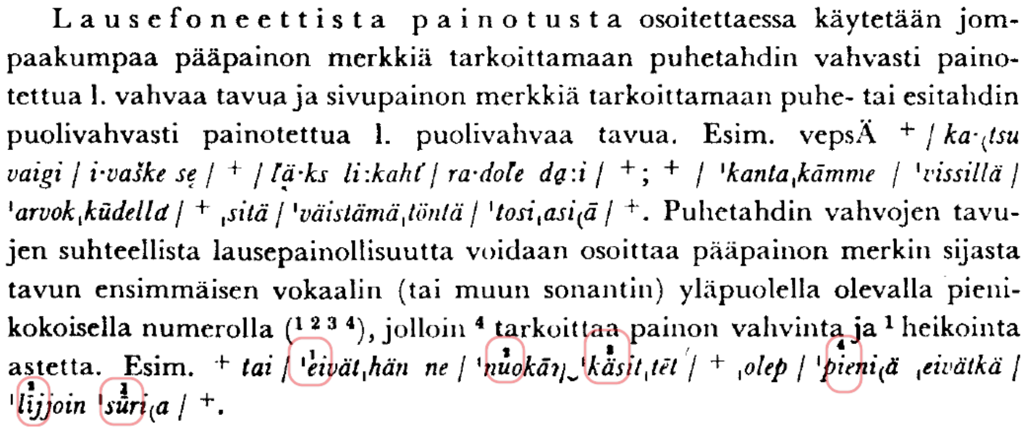
**Fig. 1968b-166:** *Showing specimens for 1, 4, 8 (red) and 2, 7 (green) applied to metrical symbols*. This specimen shows a superscript number larger than 9, formed by two superscript digits each placed side by side, thus overriding the default stacking behavior otherwise applied to sequences of diacritical marks with the same combining class.



metrum: dactyloepitr. Α′–Γ'

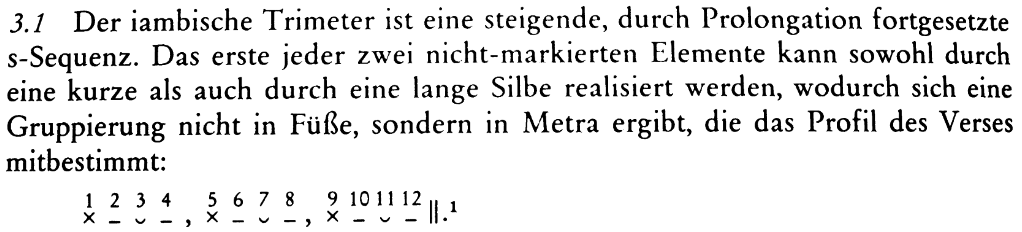
ΣΤΡ
— D ◡ ⋮ 6 E || 2 E e — D | 3 E ◡ D × |
4 D ◡ e × | 5 E — e — |||
ΕΠ
D | — D ◡ e || 2 D — e — || 3 E — ||
4 E — d1 || 5 e — D | 6 e ◡◡ — e — |||
1
8
4
27

**Fig. 1977a-7:** *Showing specimens for 1...4 on Latin letters.*



Lausefoneettista painotusta osoitettaessa käytetään jompaakumpaa pääpainon merkkiä tarkoittamaan puhetahdin vahvasti painotettua l. vahvaa tavua ja sivupainon merkkiä tarkoittamaan puhe- tai esitahdin puolivahvasti painotettua l. puolivahvaa tavua. Esim. vepsÄ + / *ka·₍tsu vaigi* / *i·vaške sę* / + / *ľä·ks li:kahť* / *ra·doľe dǫ:i* / +; + / *ˈkanta͵kāmme* / *ˈvissillä* / *ˈarvok͵küdellä* / + *͵sitä* / *ˈväistämä͵töntä* / *ˈtosi͵asi₍ā* / +. Puhetahdin vahvojen tavujen suhteellista lausepainollisuutta voidaan osoittaa pääpainon merkin sijasta tavun ensimmäisen vokaalin (tai muun sonantin) yläpuolella olevalla pienikokoisella numerolla (1 2 3 4), jolloin 4 tarkoittaa painon vahvinta ja 1 heikointa astetta. Esim. + *tai* / *ˈeivät͵hän ne* / *ˈnuokāŋ‿ˈkäsit͵tēt* / + *͵olep* / *ˈpieni₍ä ͵eivätkä* / *ˈlijjoin ˈsüri₍a* / +.

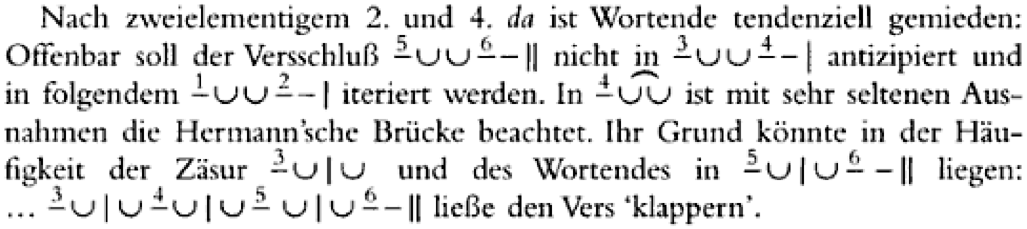
**Fig. 1993a-88:** *Showing specimens for 0...9, applied to metrical symbols*. This specimen shows superscript larger than 9, formed by two superscript digits each placed side by side, thus overriding the default stacking behavior otherwise applied to sequences of diacritical marks with the same combining class.



*3.1* Der iambische Trimeter ist eine steigende, durch Prolongation fortgesetzte s-Sequenz. Das erste jeder zwei nicht-markierten Elemente kann sowohl durch eine kurze als auch durch eine lange Silbe realisiert werden, wodurch sich eine Gruppierung nicht in Füße, sondern in Metra ergibt, die das Profil des Verses mitbestimmt:

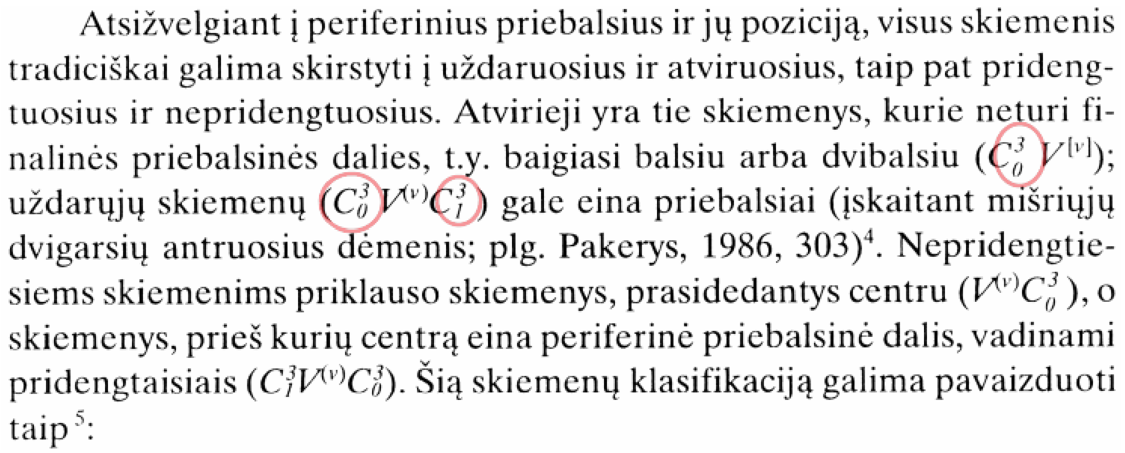
1 2 3 4 5 6 7 8 9 10 11 12
× — ◡ — , × — ◡ — , × — ◡ — ||.1

**Fig. 1997a-348:** *Showing specimens for 1...6 applied to metrical symbols.*



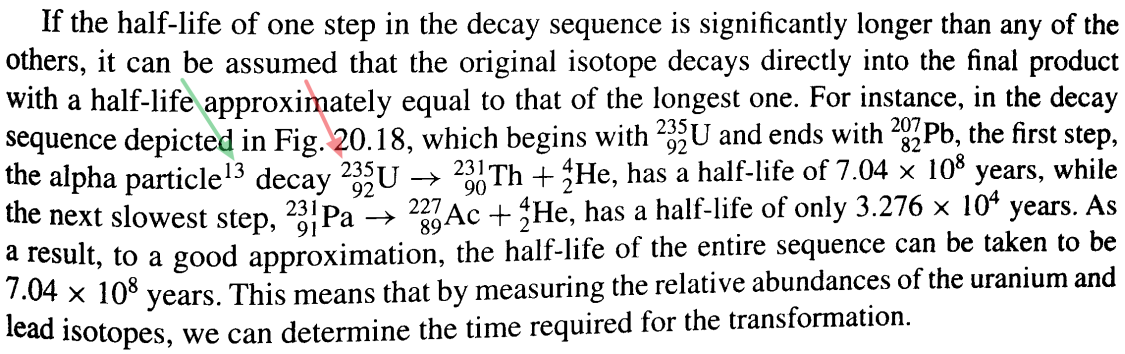
Nach zweielementigem 2. und 4. *da* ist Wortende tendenziell gemieden: Offenbar soll der Versschluß $\overset{5}{-}\cup\cup\overset{6}{-}-\|$ nicht in $\overset{3}{-}\cup\cup\overset{4}{-}-|$ antizipiert und in folgendem $\overset{1}{-}\cup\cup\overset{2}{-}-|$ iteriert werden. In $\overset{4}{-}\overset{\frown}{\cup\cup}$ ist mit sehr seltenen Ausnahmen die Hermann'sche Brücke beachtet. Ihr Grund könnte in der Häufigkeit der Zäsur $\overset{3}{-}\cup|\cup$ und des Wortendes in $\overset{5}{-}\cup|\cup\overset{6}{-}-\|$ liegen: … $\overset{3}{-}\cup|\cup\overset{4}{-}\cup|\cup\overset{5}{-}\cup|\cup\overset{6}{-}-\|$ ließe den Vers 'klappern'.

**Fig. 2001a-119:** *Showing specimens for COMBINING DIGIT THREE ABOVE applied over a subscript digit (U+2080 resp. U+2081).*



Atsižvelgiant į periferinius priebalsius ir jų poziciją, visus skiemenis tradiciškai galima skirstyti į uždaruosius ir atviruosius, taip pat pridengtuosius ir nepridengtuosius. Atvirieji yra tie skiemenys, kurie neturi finalinės priebalsinės dalies, t.y. baigiasi balsiu arba dvibalsiu ($C_{\overset{3}{0}}V^{[v]}$); uždarųjų skiemenų ($C_{\overset{3}{0}}V^{(v)}C_{\overset{3}{1}}$) gale eina priebalsiai (įskaitant mišriųjų dvigarsių antruosius dėmenis; plg. Pakerys, 1986, 303)[4]. Nepridengtiesiems skiemenims priklauso skiemenys, prasidedantys centru ($V^{(v)}C_{\overset{3}{0}}$), o skiemenys, prieš kurių centrą eina periferinė priebalsinė dalis, vadinami pridengtaisiais ($C_{\overset{3}{1}}V^{(v)}C_{\overset{3}{0}}$). Šią skiemenų klasifikaciją galima pavaizduoti taip[5]:

**Fig. 2007a-757:** *Showing specimens of several combining superscript digits to denote nuclides named in plain text*
*Note that the U+00B3 SUPERSCRIPT THREE (marked green) has not the same size as the "combining digits above", thus the "2" marked red has to be encoded U+00A0 NO-BREAK SPACE + COMBINING DIGIT TWO ABOVE to be in line with the other "combining digits above" which are above subscript digits.*



If the half-life of one step in the decay sequence is significantly longer than any of the others, it can be assumed that the original isotope decays directly into the final product with a half-life approximately equal to that of the longest one. For instance, in the decay sequence depicted in Fig. 20.18, which begins with $^{235}_{92}$U and ends with $^{207}_{82}$Pb, the first step, the alpha particle[13] decay $^{235}_{92}\mathrm{U} \rightarrow {}^{231}_{90}\mathrm{Th} + {}^{4}_{2}\mathrm{He}$, has a half-life of $7.04 \times 10^8$ years, while the next slowest step, $^{231}_{91}\mathrm{Pa} \rightarrow {}^{227}_{89}\mathrm{Ac} + {}^{4}_{2}\mathrm{He}$, has a half-life of only $3.276 \times 10^4$ years. As a result, to a good approximation, the half-life of the entire sequence can be taken to be $7.04 \times 10^8$ years. This means that by measuring the relative abundances of the uranium and lead isotopes, we can determine the time required for the transformation.

**ISO/IEC JTC 1/SC 2/WG 2**
**PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS**
**FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646**[1]
**Please fill all the sections A, B and C below.**
**Please read Principles and Procedures Document (P & P) from http://www.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://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.**
**See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html for latest *Roadmaps*.**

## A. Administrative

1. **Title:** ***Proposal to encode combining decimal digits above in the UCS***
2. Requester's name: *Martin Schrage; Karl Pentzlin*
3. Requester type (Member body/Liaison/Individual contribution): *Expert Contribution*
4. Submission date: *2011-10-15*
5. Requester's reference (if applicable): *University of Munich, Germany (M. S.)*
6. Choose one of the following:
This is a complete proposal: *Yes*
(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): *No*
Proposed name of script:
b. The proposal is for addition of character(s) to an existing block: *Yes*
Name of the existing block: *Combining Diacritical Marks Extended*
2. Number of characters in proposal: *10*
3. Proposed category (select one from below - see section 2.2 of P&P document):
A-Contemporary *X* 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? *Yes*
a. If YES, are the names in accordance with the “character naming guidelines” in Annex L of P&P document? *Yes*
b. Are the character shapes attached in a legible form suitable for review? *Yes*
5. Fonts related:
a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
*The authors (if requested)*
b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
*The authors (if requested)*
6. References:
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *Yes*
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *Yes*
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)? *Yes*
*See text*

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 http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N3702-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)

**C. Technical - Justification**

1. Has this proposal for addition of character(s) been submitted before? *Yes*
If YES explain *They are contained in WG2 N3913 = L2/10-358R and are separated here from its revision*

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? *One of the authors (M. S.) is a member of the scientific community himself*
If YES, available relevant documents: *See text*

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *Yes*
Reference: *See text*

4. The context of use for the proposed characters (type of use; common or rare) *Common*
Reference: *See text*

5. Are the proposed characters in current use by the user community? *Yes*
If YES, where? Reference: *See text*

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *Yes*
If YES, is a rationale provided? *Yes*
If YES, reference: *To keep them in line with related characters*

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 an existing character? *Yes*
If YES, is a rationale for its inclusion provided? *Yes*
If YES, reference: *See text*

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? *Yes*
If YES, reference: *See text*
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *n/a*
If YES, reference: *The proposal contains combining characters but no composite sequences*

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 character(s)? *No*
If YES, is the equivalent corresponding unified ideographic character(s) identified?
If YES, reference: