Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Responses to NCITS/L2 and Unicode Consortium comments on numerous proposals**
Source: **Michael Everson**
Status: **Working Document**
Date: **1998-09-12**

This document responds to L2/98-293 and L2/98-292, which comment on proposals for new characters to be added to the UCS (SC2/WG2 N1741, N1742, N1743, N1744, N1745, N1746, N1747, N1748, N1749).

> L2/98-293
>
> 2.1. N 1741: Additional Latin characters for the UCS
>
> 0222 LATIN CAPITAL LETTER Z WITH HOOK
> 0223 LATIN SMALL LETTER Z WITH HOOK
>
> NCITS/L2 and the Unicode Consortium recommend addition of these two characters to ISO/IEC 10646 with the proposed names and code positions.

I am delighted to hear this.

> 0224 LATIN CAPITAL LETTER OU
> 0225 LATIN SMALL LETTER OU
>
> NCITS/L2 and the Unicode Consortium have grave concerns about the cloning of letters that are clearly derived from another script.

In the case of these characters, used in the Algonquin language in Canada, while it is true that a French Catholic priest introduced the characters and derived them from the Greek ligatures with which he was familiar (one *infers* this from the glyphs; it is not recorded anywhere), they have been *assimilated and naturalized* into the Latin script. Indeed, while may both be the LATIN LETTER OU and the GREEK LETTER OU are both written as a kind of twisted loop open at the top, the Algonquin letters are usually closed at the top, and with typewriter technology the digit 8 has often been substituted for them. (indeed the glyphs proposed for the LATIN LETTER OU might be revised to emphasize this). The pictures given here are from *Le Noël huron* (1648), a roadsign (1998), and a school (1998.

Scripts do borrow letters from one another, and do assimilate and naturalize those letters into the orthographies of natural languages, and into other conventional systems such as the IPA. An example is the LATIN LETTER THORN, borrowed from the Runic script more than a thousand years ago. LATIN LETTER OU is another, albeit borrowed “only” 350 years ago.

2.2. N 1742: Additional IPA characters for the UCS

NCITS/L2 and the Unicode Consortium recommend preparation of a new proposal for IPA characters with the following contents:

- The first four characters in this proposal
- The last two characters (034F and 0350) from N 1746

I have made such a proposal (N1845).

The last three characters in this proposal (02AD through 02AF) which are IPA clones of Greek letters should not be proposed until the issue of script-specific clones of characters is resolved.

Further discussion on the LATIN LETTER BETA, LATIN LETTER THETA, and LATIN LETTER CHI will be taken up in another document.

2.3. N 1743: Additional Greek characters for the UCS

03D7 GREEK KAI SYMBOL
03F4 GREEK CAPITAL LETTER OU
03F5 GREEK SMALL LETTER OU

NCITS/L2 and the Unicode Consortium recommend addition of these three characters to ISO/IEC 10646 with the proposed names and code positions.

I am delighted to hear this. I hope that ΕΛΟΤ will also favour the addition of these characters.

03D8 GREEK KAI SYMBOL WITH VARIA

NCITS/L2 and the Unicode Consortium consider that there is no need to add this character. Once the proposed GREEK KAI SYMBOL has been added, the GREEK KAI SYMBOL WITH VARIA would then also be encoded using existing ISO/IEC 10646 characters. The argument that all Greek characters must be implementable at Level 1 is not convincing when dealing with characters such as this that derive from manuscript traditions, rather than with general modern-use Greek characters.

It is true that GREEK KAI SYMBOL WITH VARIA could be composed in a Level 3 implementation. I did not see the precomposed symbol in use in modern Athenian signage, though of course it, and many other ligatures, are common in early Greek typography. The argument that all Greek characters must be implementable at Level 1 is weak in the case of this character particularly in view of the fact the the argument *for* including the base character is based on the *modern* usage of the character in signage, not its usage as one of many ligatures in early Greek typography.

2.4. N 1744: Additional Cyrillic characters for the UCS

051A CYRILLIC CAPITAL LETTER EL WITH DESCENDER
through
0523 CYRILLIC SMALL LETTER E WITH DIAERESIS

051A CYRILLIC CAPITAL LETTER EL WITH DESCENDER
051C CYRILLIC CAPITAL LETTER ER WITH TICK
051E CYRILLIC CAPITAL LETTER SHORT I WITH DESCENDER
0520 CYRILLIC CAPITAL LETTER EM WITH DESCENDER
0522 CYRILLIC CAPITAL LETTER E WITH DIAERESIS

Only the lower case forms of these characters are found in the Kildin Sámi alphabet (as documented by the Sámi Sector of the Murmansk Branch of the Russian Academy of Science, WG2 number for this document not yet assigned). NCITS/L2 and the Unicode Consortium consider that these upper case characters should not be added to ISO/IEC 10646 because there is insufficient evidence to document their use.

This is a rather disappointing interpretation, and seems almost to willfully ignore the simplest of principles of the Cyrillic (and Greek and Latin) scripts, namely that *letters have case*. It is only exceptional cases, such as LATIN LETTER BILABIAL CLICK and CYRILLIC LETTER PALOCHKA, where

letters in these scripts do not have case. Trond Trosterud would certainly not have proposed capital letters had they not existed (that is, had they been analogous to the BILABIAL CLICK and the PALOCHKA). Rather than formulating a consideration that "they should not be added to ISO/IEC 10646 because there is insufficient evidence to document their use", it would be more helpful if NCITS/L2 and the Unicode Consortium were simply to request confirmation that casing forms of these characters are in fact required.

The alphabet given in the fax sent by the Sámi Sector of the Murmansk Branch of the Russian Academy of Science is identical to the alphabet given on page 16 and page 529 of R. D. Kuruč's Са̄мь-рӯшш соагкнэнкь (Саамско-русский словарь) (Москва: Русский язык 1985). This alphabet is considered authoritative, which is why Kuruč and Afanas´eva reproduced it in their fax. However, it was never intended for the use of character set standardizers, and therefore, by omitting the capital forms of the letters in question, provides the (to us) totally irrelevant information that the characters in question do not occur word-initially («Буквы ӈ, ӎ, ӆ, ҏ, ӈ, j, h, ӓ, ӛ в начале слов не употребляются» 'The letters ӈ, ӎ, ӆ, ҏ, ӈ, j, h, ӓ, ӛ do not appear at the beginning of words'). Throughout the dictionary, however, headwords are given in all capitals, and these letters occur there.

> 0523 CYRILLIC SMALL LETTER E WITH DIAERESIS This character is part of the Kildin Sámi alphabet. NCITS/L2 and the Unicode Consortium consider that there is no need to add this character as it is already encoded with existing ISO/IEC 10646 characters.

I disagree with this assessment; other Kildin Sámi letters (Ä ä Ë ë) are available as unitary characters in the UCS.

> NCITS/L2 and the Unicode Consortium recommend addition of the remaining four characters with the recommended names and code positions.

I recommend accepting all ten characters.

> 2.5. N 1746: Additional combining characters for the UCS
> See N 1742: Additional IPA characters for the UCS above.

See N1845.

> L2/98-292
>
> 2.1. N 1741: Additional Latin characters for the UCS
>
> 0220 LATIN CAPITAL LETTER KRA
>
> No proof of the existence of an uppercase form of kra as an independent character in actual use has been presented.

KRA was used in Kleinschmidt's orthography of Greenlandic, introduced in 1851. This orthography was changed in 1973. Personal names, placenames, and sentence initials and words in an all-caps context now written with a CAPITAL LETTER Q were formerly written with a CAPITAL LETTER KRA (i.e., Qaanaaq 'Thule' would have been written something like K'ãnaoĸ (I am not sure of the vowels).

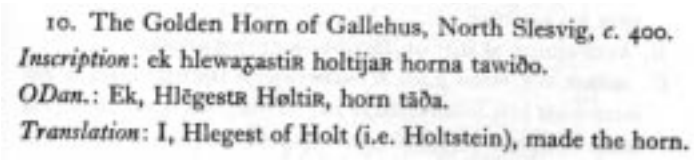
> 0221 LATIN CAPITAL LETTER YR
>
> This character is already in ISO/IEC 10646:1 as LATIN LETTER YR, encoded at position U+01A6. LATIN SMALL LETTER YR (encoded at position 79 in ISO 5426-2:1996) has been misidentified as U+01A6 LATIN LETTER YR.

We assumed that LATIN LETTER YR was lower-case just as LATIN LETTER WYNN was. The letter YR is used to represent a retroflex *z* of sorts, deriving from an original Germanic *-z*, ultimately deriving from an Indo-European *-s*. By about 1000 the sound had fallen together with *r* as in modern Icelandic. It is syllable final. Which means if you think that LATIN LETTER YR is a capital letter we're

in big trouble because Old Danish texts like *Ek, HlégestR HøltiR, horn táða* ‘I, Hlegest of Holt(stein), made the horn’ can’t currently be represented in the UCS (except in Runic of course).

> Proof of the existence of a lowercase form of yr as an independent character in actual use should be provided if the lowercase form is to be proposed for addition to ISO/IEC 10646. NCITS/L2 and the Unicode Consortium note that the British Library Special character set which was a primary source for ISO 5426-2:1996 is consistent with the current repertoire of ISO/IEC 10646. The British Library set (shown in Figure 1) includes only an x-height GREENLANDIC K (code E3) and an h-height SWASH R (code EF), but not the inverse equivalents with respect to height.

If the current LATIN LETTER YR is really intended to be the capital letter (which would accord with its identity as SWASH R), then LATIN SMALL LETTER YR needs to be added. The glyph shape of the LATIN LETTER YR is *really* weird in any case. In ordinary text the YR is rather like the KRA; it is


10. The Golden Horn of Gallehus, North Slesvig, *c.* 400.
Inscription: ek hlewagastiR holtijaR horna tawiđo.
ODan.: Ek, Hlēgestr Høltir, horn tāda.
Translation: I, Hlegest of Holt (i.e. Holtstein), made the horn.

always presented as a small caps R. That is the only form I have ever seen. I have never seen, ever, – apart from in ISO/IEC 10646 and the Unicode Standard – the form with the high ascender (Ʀ looking like an I-R ligature) – and I did a fair bit of Germanic linguistics.

A capital form of YR would be rare enough, to be used in all caps environments (since it’s typically only found in final position – but this can be morpheme-final, since the definite article *-inn* may follow it, for instance). Adding the small YR would be the reciprocal of what we did with the two other Germanicist characters, HWAIR and WYNN. In any case we need both cases of this character 1) because case is a normal part of the Latin script (apart from exceptional letters like the clicks), 2) because Germanicists may well wish to use YR in an all-caps context, and 3) because the two cases are distinguished in the TC46 ISO standard.

> 2.2. N 1743: Additional Greek characters for the UCS
>
> 03DB GREEK SMALL LETTER STIGMA
> 03DD GREEK SMALL LETTER DIGAMMA
> 03DF GREEK SMALL LETTER KOPPA
> 03E1 GREEK SMALL LETTER SAMPI
>
> The character repertoire of Version 1.0 of The Unicode Standard included these characters. Examples of use have been included in N 1743 for all except small letter sampi. These characters should be added to ISO/IEC 10646 with the recommended code values and names.

I am delighted to hear this.

> 2.3. N 1744: Additional Cyrillic characters for the UCS
> Numeric Signs
>
> 0487 CYRILLIC TEN THOUSANDS SIGN
> 0488 CYRILLIC HUNDRED THOUSANDS SIGN
> 0489 CYRILLIC MILLIONS SIGN
>
> NCITS/L2 and the Unicode Consortium recommend:
> • Unification of CYRILLIC TEN THOUSANDS SIGN with U+20DD, COMBINING ENCLOSING CIRCLE;
> • Acceptance of the remaining two characters with the recommended names and code positions;

I am delighted to hear this, but would like to raise the question as to whether the generic COMBINING ENCLOSING CIRCLE can be used unambiguously with numeric value in implementations of this kind. If it cannot, then the CYRILLIC TEN THOUSANDS SIGN should be added to the UCS. If the unification is accepted, the 2 characters should be moved up one position each.

> Cyrillic Characters for non-Slavic Languages

04C5 CYRILLIC CAPITAL LETTER CHECHEN KA
through
0519 CYRILLIC SMALL LETTER KOMI TJE.

NCITS/L2 and the Unicode Consortium oppose addition of the characters in the following ranges to ISO/IEC 10646 because they duplicate existing characters: 04C5-04C6, 04C9-04CA, 04CD-04CE, 04EC-04ED, 04F6-04F7, 04FA-04FD, 0508-0515. For details, see Appendix.

I disagree strongly with the assertion that these characters are duplicates of other characters in the UCS, but agree that further study on all of these characters is warranted.

The Unicode Technical Committee and NCITS/L2 consider that the remaining letters (04FE-04FF, 0500-0507, 0516-0519) are doubtful and should not be added to ISO/IEC 10646 at this time. Further study is needed to determine whether they are truly unique or are glyphic variants of existing letters. The first pair (04FE-04FF) are from the Missionary orthography of the Mordvin-Moksha Dialect and could be the case forms of a typographic digraph; the others are all from the 1919 (Molodtsov) orthography for the Komi language.

It should be noted that the Molodtsov orthography for Komi uses a right upturn glyph regularly to make a new letter. Unifying some of these characters with characters whose modification is based not upon such an upturn glyph, but on the SOFT SIGN, would lead to a situation where either Komi text or Serbian text could be presented with totally inauthentic glyphs. The kind of glyph unification adduced by NCITS/L2 and the Unicode Consortium does not match the normal Cyrillic practice, and frankly, I consider the unifications in the *ALA Romanization Tables* to be influenced, practically, as an illustration to library cataloguers, by the diacritical marks hand-drawn over typescript Cyrillic found in Soviet books printed on *bumaga* № 3 than by the realities of any actual glyph identities recognizable by the users. I agree that further study on these characters is warranted.

2.4. N 1745: Additional Math characters for the UCS

22F2 VECTOR OR SUM
22F3 VECTOR PRODUCT
22F4 SUM OR UNION OF CLASSES OR SETS
22F5 PRODUCT OF INTERSECTION OF CLASSES OR SETS
22F6 IS INCLUDED IN SET
22F7 INCLUDES IN SET

These characters are from Table 2, Extension of Basic Set G0 of ISO 6862:1996. ISO DIS 6862 was a source for Version 1.0 of The Unicode Standard. In the opinion of the Unicode Working Group (predecessor to the Unicode Technical Committee), these characters replicated characters in Table 1, Basic Set G0. The Unicode Technical Committee re-examined the characters and came to the same conclusion. NCITS/L2 concurs. NCITS/L2 and the Unicode Consortium oppose addition of this whole collection to ISO/IEC 10646 because they duplicate existing characters.

Randy Barry and I were unaware that such scrutiny had already been applied to these characters. For my part I accept NCITS/L2 analysis, and agree that these characters should not be added to the UCS. Roundtrip conversion between ISO 6862 and ISO/IEC 10646 will not be possible, but such operations will result in normalization of the 6862 text by abandonment of the replicated characters.

2.5. N 1746: Additional combining characters for the UCS

0346 COMBINING RIGHT DESCENDER
0347 COMBINING LEFT DESCENDER

These combining marks are intended for use with Cyrillic letters (as described in Clauses 6.2 and 6.4 of ISO 10754). Table A.2 in ISO 10754 documents combinations of Cyrillic letters and descenders. The ISO 10754 sequence of a combining descender followed by a letter can be mapped to extant characters in ISO/IEC 10646:1, except for one case pair which is proposed for addition in N 1744. The combining descenders are therefore redundant. NCITS/L2 and the Unicode Consortium also oppose addition of these characters to ISO/IEC 10646, because of the serious implications their addition would have for decomposition, which:
• will impact the development and implementation of other standards (such as the International String Ordering standard, 14651); and,
• can destabilize the effort underway to define standard normalization forms of Unicode/10646 for use by W3C and by the programming languages community.

These reasons for not accepting these characters are good. Mapping of ISO 10754 data which contains these characters and a base form should be many-to-one to the UCS and one-to-many from the UCS.

> Combining Small Letters Above
>
> 0348 COMBINING LATIN SMALL LETTER A ABOVE
> 0349 COMBINING LATIN SMALL LETTER E ABOVE
> 034A COMBINING LATIN SMALL LETTER R ABOVE
> 034B COMBINING LATIN SMALL LETTER Z ABOVE
>
> The only documentation for these characters is ISO 5426-2:1996 itself.
> The proposed character COMBINING LATIN SMALL LETTER E ABOVE is an early form of the umlaut. US cataloging practice is to substitute the umlaut (Descriptive Cataloging of Rare Books, p. 69).

Umlaut? Diaeresis? We should endeavour to get further information on these characters from TC46/SC4/WG1.

> 034C COMBINING DOUBLE CARON
> 034D COMBINING DOUBLE CIRCUMFLEX
> 034E COMBINING CIRCUMFLEX GRAVE
>
> The only documentation for these characters is ISO 5426-2:1996 itself. These marks appear to be Latin contractions (see comments on N 1747 below).

The CIRCUMFLEX GRAVE is an IPA character (see N1845) I will endeavour to get further information on the other characters from TC46/SC4/WG1. The name should definitely be GRAVE CIRCUMFLEX, not the reverse.

> 2.6. N 1747: Contraction characters for the UCS
>
> 00 LATIN CONTRACTION AGUS (et, ond)
> through
> 15 LATIN SMALL CONTRACTION LONG S WITH HOOK
> also, 2048 REVERSED SECTION SIGN in N 1748, and possibly many of the combining characters in N 1746.
>
> LATIN CONTRACTION AGUS
> Mr. Michael Everson has said that this character is the modern Irish equivalent of the ampersand. The British Library's "Special" character set from which ISO 5426-2 is partially derived calls this character "Ampersand form type 2." Descriptive Cataloging of Rare Books (2nd. ed., Library of Congress, 1991) instructs the cataloger (p. 69): If the Tironian sign (<image of Tironian sign>) cannot be reproduced, treat it as an abbreviation and substitute "[et]" for it.
>
> This existence of this character is documented by three sources. The US rules for rare book cataloging prescribe its use. It should be added to ISO/IEC 10646, but as a General Punctuation character with the proposed name TIRONIAN SIGN.

Without prejudice to the possibility of including the other contractions, it has to be said that the name of this character still needs to be settled. Calling it TIRONIAN SIGN is not sufficient – the LC's *Descriptive catalogue of rare books* names the character not as "Tironian sign" but as "Tironian sign ⁊". There were lots of Tironian signs (see appendix A from Faulmann). N1747 differentiated between the sign ⁊ and ⁊ because ISO 5426-2:1996 did so, and we called the former LATIN CONTRACTION ET because it was used in French and the latter LATIN CONTRACTION AGUS because this character is in common use today in Irish Gaelic to represent the word *agus* 'and' (it is also often used in the form ⁊rl 'etc'. We noted that the sign can also be used for Latin and French *et* 'and' and Old English *ond* 'and'. TIRONIAN SIGN ET would be an acceptable name for this character, and should the French one be accepted, it should then be named TIRONIAN SIGN ET WITH STROKE.

> Mappable Latin Contractions
>
> 06 LATIN CONTRACTION REVERSED US

07 LATIN CONTRACTION IS
08 LATIN CONTRACTION SMALL IS
09 LATIN CONTRACTION UM

These characters can be mapped to existing UCS characters and so should be eliminated from the proposal.

The mappings provided in L2/98-292 are specious, though the glyphs in ISO 5426-2 would support the unifications. Cappelli's *Dizionario di abbreviature latine ed italiane* however shows that the glyphs, despite their names in ISO 5426-2, are different enough to make these unifications inappropriate. If, for reasons of compatibility, the contractions should be added to the UCS, these four should be among them.

General Comments about Latin Contractions
US practice for rare book cataloging is to substitute the spelled-out equivalent for the contraction. Rule 0J2 of Descriptive Cataloging of Rare Books (p. 7) states:

> When special marks of contraction have been used by the printer in continuance of the manuscript tradition, expand affected words to their full form and enclose supplied letters in square brackets. When an abbreviation standing for an entire word appears in the source, record instead the word itself, and enclose it in square brackets.

(The exception to this rule, noted above, is the Tironian sign.)

The US position is that these contraction signs should not be added to ISO/IEC 10646. Contraction signs of this general sort might be useful additions as glyphic forms for fonts used in the facsimile reproduction of manuscripts and early printed books, but their encoding as characters for the representation of textual content is counterproductive. Their use for the representation of text would cause searching and comparison problems in electronic texts and cataloging records so encoded.

Furthermore, the Latin contractions in ISO 5426-2 are only a small subset of the contractions used over the centuries in manuscripts and later included in some early printed works. Contractions are not exclusive to Latin manuscripts. Examples of Cyrillic contractions appear in the documentation accompanying N 1744.

Untrue. Some of the characters in N1744 *may* be ligatures; none of them are contractions.

here is no good rationale to merely pick out the small subset of such manuscript contraction forms present in ISO 5426-2 and propose them for encoding, as opposed to any other set which could be brought forward. While the Unicode Consortium and NCITS/L2 do not recommend the encoding of any of these contraction signs (other than the agus/Tironian sign) in ISO/IEC 10646, if WG2 chooses to accept any of the remaining characters proposed in N 1747 (i.e., 01-15 and 0A-15), then they should:

a) be considered Letterlike Symbols, for use in any script context; and,
b) be given descriptive names (e.g., those in ISO 5426-2:1996) rather than names based on the meaning of the contraction sign in a particular language context.

The names proposed in N 1747 are generally unacceptable.

I disagree. If they are added, they should be added as what they *are* – Latin contractions. It is true – a quick look at Cappelli shows this clearly – that there are thousands of contractions. These few characters should be added to the UCS as compatibility characters *if* they have been used in data. We welcome L2's advice on how to determine if this has been the case. Apparently US bibliographical practice is not to use them, but perhaps they have been used in the United Kingdom or elsewhere?

The compatibility argument is compelling here – if anyone has expanded contractions in a book to full Latin forms in square brackets in a bibliographical record, then the characters can be abandoned because they weren't used. But if they have been used in data, then we have the problem of the polyvalence of the characters themselves – since they could be *-is* or *-us* or *-arum* or *-orum,* etc., it would be impossible to map them.

We (Randy Barry and I) don't want to add spurious characters to the UCS. This spate of proposals has been geared to support of the TC46/SC4 standards as part of TC46/SC4's abandonment of development of character sets in favour of the authority of JTC1/SC2. We earnestly request L2's assistance in helping us determine whether these contraction characters have been used or not.

2.7. N 1748: Additional signature mark characters for the UCS

2048 REVERSED SECTION SIGN

This character is a Latin contraction, not a signature mark.
The image of this character is incorrect (as is its name, which is based on the incorrect image). ISO 5426-2:1996 says that this is ‘Used for the Latin suffix “orum”.’ The British Library’s “Special” character set includes a character “-RUM WORD ENDING TYPE 2” which may be the origin of this character. If characters from ISO 5426-2 representing Latin contractions are added to ISO/IEC 10646:1, this character should not be added until it has been properly identified.

I agree. I can’t find this character in Cappelli either.

2614 SIX-SPOKED ASTERISK

CHASE is an EU-funded project to develop Unicode/UCS mappings for character sets used in European libraries. CHASE has mapped the British Library’s equivalent of this character to U+2736, SIX POINTED BLACK STAR. This character should be eliminated from the proposal.

I agree. The signature mark can easily be mapped to this character, because basically signature marks are dingbats. However, in researching this I did find a reference which reminded me of a character which is certainly missing in the UCS, in Carl Liungman's *Tanketecken*. In astrology, a 60-degree angle between two planets is called its sextile aspect. The symbol for SEXTILE is invariably drawn (I have it here in Liungman, published in Trondheim in 1993, and in the American Heritage Dictionary published in Boston in 1981) as a horizontal bar imposed over an x, with nice straight lines. The Unicode Standard has unified SEXTILE with the Zapf Dingbat U+2736 SIX POINTED BLACK STAR. This unification is, I believe, incorrect, because (as a Zapf Dingbat) the dingbat doesn't allow shape variation and it is a vertical bar imposed over an x, and it does not have straight lines, but is star shaped. A star is not an asterisk. I am certain that astrological fonts will not ever use a character that looks like U+2736.

A new character SEXTILE should be added to the UCS.

The following characters should be added to ISO/IEC 10646:1, each with the proposed name and code value:

2049 REVERSED PILCROW SIGN
2139 LATIN CAPITAL LETTER ROTATED Q

The image of this character should more closely resemble to the British Library source character.

I am delighted to hear it. Of course the glyph can be changed, though I note that the British Library source character differs from the one in the ISO standard. One presumes that a printer simply took a Q sort and turned it.

2183 ROMAN NUMERAL REVERSED ONE HUNDRED

This character is not a section symbol but a Roman numeral. Mr. Michael Everson provided convincing examples of use.

I am delighted to hear it.

The following characters should be added to ISO/IEC 10646:1, with the proposed names but in the General Punctuation block.

2615 BLACK LEFTWARDS BULLET
2616 BLACK RIGHTWARDS BULLET
2768 REVERSED ROTATED FLORAL HEART BULLET

I am delighted to hear it. I don’t really understand the rationale for not keeping the REVERSED ROTATED FLORAL HEART BULLET next to the ROTATED FLORAL HEART BULLET, but it doesn’t matter

to me.

> General Comments about Signature Marks
> The marks used by printers to identify signatures (also called collations or gatherings) are important for the study of early printing. The repertoire of signature marks in ISO 5426-2 is only a small subset of the marks used by printers. Since ISO/IEC 10646 is intended to encode plain text, and the particular font used in a signature mark may be significant, the full and correct representation of signatures marks cannot be achieved without the use of a higher level protocol. The Unicode Consortium and NCITS/L2 note that the bullets and the reversed pilcrow sign are general-purpose punctuation marks.

I agree.

> 2.8. N 1749: Additional Hebrew cantillation characters for the UCS
>
> 05F5 HEBREW ACCENT TSERE
> through
> 05FC HEBREW ACCENT ASTERISK
>
> These proposed characters are from Table 2: Special vowel points, accents, and marks of ISO 8957:1996. The contents of Table 2 are historic cantillation marks. There are no known implementations of this table. The decision to equate certain Hebrew cantillation marks with diacritical marks (for example, 41 HEBREW ACCENT TSERE was mapped to U+0308 COMBINING DIAERESIS) is questionable

We were trying to add as few characters as possible and make as many reasonable unifications as possible. Maybe this one wasn't so reasonable.

> WG1 also mapped 5A HEBREW ACCENT RAFE to U+05BF HEBREW POINT RAFE, even though the two characters had different names and images. Furthermore, 4C HEBREW POINT RAFE in Table 1 is mapped to U+05BF HEBREW POINT RAFE, so that the net effect is to unify 4C in the Basic Hebrew alphabet and 5A in the Special vowel points, accents, and marks.

Oops.

> NCITS/L2 and the Unicode Consortium recommend that the addition of characters from Table 2 of ISO 8957:1996 be deferred. This will allow experts on cantillation marks and historic pointing and scholars from Israel and from other countries to clarify the nature of specific marks, and to determine the correct mappings for certain TC 46 characters. This will allow a complete proposal for any missing characters to be developed.

I certainly agree with this. We would like WG2 to request that SII, in consultation with other experts inside and outside Israel, indicate what the correct UCS mappings are for ISO 8957, and to indicate the status of any of the unmappable characters.

> Appendix: Characters from ISO 10754:1996

I am not going to go into detail on these Cyrillic characters here, because I believe that further research is warranted. However I will make a few notes with regard to NCITS/L2 and the Unicode Consortium's proposed unifications of some of these with existing characters.

> 04C5 CYRILLIC CAPITAL LETTER CHECHEN KA
> 04C6 CYRILLIC SMALL LETTER CHECHEN KA
>
> Already encoded. The uppercase Chechen ka is U+041A + U+030A and the lowercase Chechen ka is U+043A + U+030A. Editorial: The images at 04C5 and 04C6 in N 1744 are incorrect; the appendage is a circle, not a curve.

It is true that the extension is a ring glyph, which makes the letter all the more a base character and not a combined character. Normal combination with of U+041A and U+043A with U+030A yields Ӄ̊ ӄ̊, not Ӄ ӄ.

> 04FA CYRILLIC CAPITAL LETTER KURDISH QA
> 04FB CYRILLIC SMALL LETTER KURDISH QA
>
> Already encoded.

Note that the *ALA Romanization Tables* and ISO 10754 show the following glyph for this character: Ԛ. This is certainly not acceptable variant of Q to any user of the Latin script. However every Kurdish child living in Russia will recognize that there are indeed “foreign” letters in the Kurdish alphabet – but those are the letters Ёё Цц Ъъ Ыы Ээ Юю Яя, which are used in Russian. They will not know – because the letter has been assimilated and naturalized – that there is any difference beyond superficial glyph identity between Cyrillic Q and Latin Q, any more than there is between Cyrillic A and Latin A or Cyrillic O and Latin O.

Summary

The following characters proposed by me were endorsed by NCITS/L2 and the UTC for addition to the UCS:

0222 LATIN CAPITAL LETTER Z WITH HOOK
0223 LATIN SMALL LETTER Z WITH HOOK
03D7 GREEK KAI SYMBOL
03F4 GREEK CAPITAL LETTER OU
03F5 GREEK SMALL LETTER OU
051B CYRILLIC SMALL LETTER EL WITH DESCENDER
051D CYRILLIC SMALL LETTER ER WITH TICK
051F CYRILLIC SMALL LETTER SHORT I WITH DESCENDER
0521 CYRILLIC SMALL LETTER EM WITH DESCENDER
03DB GREEK SMALL LETTER STIGMA
03DD GREEK SMALL LETTER DIGAMMA
03DF GREEK SMALL LETTER KOPPA
03E1 GREEK SMALL LETTER SAMPI
0488 CYRILLIC HUNDRED THOUSANDS SIGN
0489 CYRILLIC MILLIONS SIGN
20xx TIRONIAN SIGN ET
2049 REVERSED PILCROW SIGN
2139 LATIN CAPITAL LETTER ROTATED Q
2183 ROMAN NUMERAL REVERSED ONE HUNDRED
2615 BLACK LEFTWARDS BULLET
2616 BLACK RIGHTWARDS BULLET
2768 REVERSED ROTATED FLORAL HEART BULLET

The following characters were not endorsed by NCITS/L2 and the UTC pending further study, or were rejected by them but I have given arguments here which I feel they may consider persuasive.

0224 LATIN CAPITAL LETTER OU
0225 LATIN SMALL LETTER OU
051A CYRILLIC CAPITAL LETTER EL WITH DESCENDER
051C CYRILLIC CAPITAL LETTER ER WITH TICK
051E CYRILLIC CAPITAL LETTER SHORT I WITH DESCENDER
0520 CYRILLIC CAPITAL LETTER EM WITH DESCENDER
0220 LATIN CAPITAL LETTER KRA
034E COMBINING GRAVE CIRCUMFLEX (see N1845)

The following characters were not endorsed by NCITS/L2 and the UTC and I disagree with their assessment (I think they should be added to the UCS).

0522 CYRILLIC CAPITAL LETTER E WITH DIAERESIS
0523 CYRILLIC SMALL LETTER E WITH DIAERESIS

The following characters were not endorsed by NCITS/L2 and the UTC and I agree with their assessment (I think they shouldn’t be added to the UCS).

03D8 GREEK KAI SYMBOL WITH VARIA
0221 LATIN CAPITAL LETTER YR
22F2 VECTOR OR SUM
22F3 VECTOR PRODUCT
22F4 SUM OR UNION OF CLASSES OR SETS
22F5 PRODUCT OF INTERSECTION OF CLASSES OR SETS
22F6 IS INCLUDED IN SET

22F7 INCLUDES IN SET
0346 COMBINING RIGHT DESCENDER
0347 COMBINING LEFT DESCENDER
2614 SIX-SPOKED ASTERISK

The following characters, arising from this discussion, should be added to the UCS.

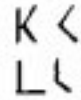
0221 LATIN SMALL LETTER YR
0221 SEXTILE

The following characters need further study.

0487 CYRILLIC TEN THOUSANDS SIGN
0348 COMBINING LATIN SMALL LETTER A ABOVE
0349 COMBINING LATIN SMALL LETTER E ABOVE
034A COMBINING LATIN SMALL LETTER R ABOVE
034B COMBINING LATIN SMALL LETTER Z ABOVE
034C COMBINING DOUBLE CARON
034D COMBINING DOUBLE CIRCUMFLEX
xx06 LATIN CONTRACTION REVERSED US
xx07 LATIN CONTRACTION IS
xx08 LATIN CONTRACTION SMALL IS
xx09 LATIN CONTRACTION UM
2048 REVERSED SECTION SIGN
05F5 HEBREW ACCENT ACUTE TSERE
05F6 HEBREW ACCENT GRAVE TSERE
05F7 HEBREW ACCENT SAMARIAN QAMATS
05F8 HEBREW ACCENT DAGESH
05F9 HEBREW ACCENT BABYLONIAN PATAH
05FA HEBREW ACCENT BABYLONIAN QAMATS
05FB HEBREW ACCENT BABYLONIAN DAGESH
05FC HEBREW ACCENT ASTERISK
04C5 CYRILLIC CAPITAL LETTER CHECHEN KA
04C6 CYRILLIC SMALL LETTER CHECHEN KA
04C9 CYRILLIC CAPITAL LETTER CHUVASH NG
04CA CYRILLIC SMALL LETTER CHUVASH NG
04CD CYRILLIC CAPITAL LETTER KOMI NG
04CE CYRILLIC SMALL LETTER KOMI NG
04EC CYRILLIC CAPITAL LETTER SELKUP OE
04ED CYRILLIC SMALL LETTER SELKUP OE
04F6 CYRILLIC CAPITAL LETTER AISOR EL
04F7 CYRILLIC SMALL LETTER AISOR EL
04FA CYRILLIC CAPITAL LETTER KURDISH QA
04FB CYRILLIC SMALL LETTER KURDISH QA
04FC CYRILLIC CAPITAL LETTER KURDISH WE
04FD CYRILLIC SMALL LETTER KURDISH WE
04FE CYRILLIC CAPITAL LETTER YA IE
04FF CYRILLIC SMALL LETTER YA IE
0500 CYRILLIC CAPITAL LETTER KOMI DE
0501 CYRILLIC SMALL LETTER KOMI DE
0502 CYRILLIC CAPITAL LETTER KOMI DJE
0503 CYRILLIC SMALL LETTER KOMI DJE
0504 CYRILLIC CAPITAL LETTER KOMI DZE
0505 CYRILLIC SMALL LETTER KOMI DZE
0506 CYRILLIC CAPITAL LETTER KOMI ZJE
0507 CYRILLIC SMALL LETTER KOMI ZJE
0508 CYRILLIC CAPITAL LETTER YAKUT I WITH STROKE
0509 CYRILLIC SMALL LETTER YAKUT I WITH STROKE
050A CYRILLIC CAPITAL LETTER JE WITH STROKE
050B CYRILLIC SMALL LETTER JE WITH STROKE
050C CYRILLIC CAPITAL LETTER KOMI ELJ
050D CYRILLIC SMALL LETTER KOMI ELJ
050E CYRILLIC CAPITAL LETTER EL WITH MIDDLE HOOK
050F CYRILLIC SMALL LETTER EL WITH MIDDLE HOOK
0510 CYRILLIC CAPITAL LETTER MORDVIN EL KA
0511 CYRILLIC SMALL LETTER MORDVIN EL KA

0512 CYRILLIC CAPITAL LETTER EN WITH MIDDLE HOOK
0513 CYRILLIC SMALL LETTER EN WITH MIDDLE HOOK
0514 CYRILLIC CAPITAL LETTER ER KA
0515 CYRILLIC SMALL LETTER ER KA
0516 CYRILLIC CAPITAL LETTER KOMI ESJ
0517 CYRILLIC SMALL LETTER KOMI ESJ
0518 CYRILLIC CAPITAL LETTER KOMI TJE
0519 CYRILLIC SMALL LETTER KOMI TJE
051A CYRILLIC CAPITAL LETTER EL WITH DESCENDER
051B CYRILLIC SMALL LETTER EL WITH DESCENDER
051C CYRILLIC CAPITAL LETTER ER WITH TICK
051D CYRILLIC SMALL LETTER ER WITH TICK
051E CYRILLIC CAPITAL LETTER SHORT I WITH DESCENDER
051F CYRILLIC SMALL LETTER SHORT I WITH DESCENDER
0520 CYRILLIC CAPITAL LETTER EM WITH DESCENDER
0521 CYRILLIC SMALL LETTER EM WITH DESCENDER
0522 CYRILLIC CAPITAL LETTER E WITH DIAERESIS
0523 CYRILLIC SMALL LETTER E WITH DIAERESIS

TIRONISCHE NOTEN.

Zeichen	Wert	Zeichen	Wert
	a		n
	b		o
	c		p
	d		q
	e		r
	f		s
	g		t
	h		u
	i		x
	k		ph
	l		te
	m		

Präfixe.

Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert
	al		ap		ar		in		prae
	ac		ad		de		inter		pro
	an		con		di		ob		re
	ab		circa		ex		per		sub

Suffixe.

Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert
	a		ans		are		atis		enus
	ae		ant		ari		atur		ens
	am		antes		aris		e		ent
	amini		anticis		arum		el		ente
	amur		antur		as		em		entes
	amus		anus		at		emur		enti

TIRONISCHE NOTEN.

Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert
⌐	entis	⌐	itur	⌐	de	⌐	enim	⌐	tat
⌐	entibus	⌐	ius	⌐	do	⌐	ra	⌐	tato
⌐	entiam	⌐	ium	⌐	dorum	⌐	rum	⌐	tatis
⌐	er	⌐	o	⌐	dum	⌐	rant	⌐	te
⌐	ere	⌐	os	⌐	la	⌐	re	⌐	ter
⌐	eris	⌐	or	⌐	lia	⌐	ret	⌐	ti
⌐	es	⌐	orem	⌐	liam	⌐	rer	⌐	tia
⌐	et	⌐	orum	⌐	lium	⌐	rere	⌐	tio
⌐	etur	⌐	os	⌐	lii	⌐	remus	⌐	tis
⌐	i	⌐	u	⌐	lis	⌐	res	⌐	tatem
⌐	ia	⌐	mus	⌐	ni	⌐	ri	⌐	to
⌐	iae	⌐	um	⌐	mam	⌐	ris	⌐	toribus
⌐	ie	⌐	unt	⌐	na	⌐	rorum	⌐	torum
⌐	ii	⌐	ur	⌐	me	⌐	s	⌐	to
⌐	iis	⌐	us	⌐	mo	⌐	sa	⌐	tos
⌐	iit	⌐	vum	⌐	mus	⌐	sem	⌐	tu
⌐	im	⌐	milia	⌐	na	⌐	sum	⌐	tum
⌐	inus	⌐	us	⌐	ne	⌐	t	⌐	tus
⌐	inus	⌐	milibus	⌐	nes	⌐	ta		
⌐	is	⌐	ci	⌐	ni	⌐	tam		
⌐	it	⌐	cum	⌐	nis	⌐	tas		

Die tironischen Noten wurden von Marcus Tullius Cicero, einem Freigelassenen des Cicero, erfunden, um dessen Reden aufzuzeichnen. Die von Tiro aufgestellten Abkürzungen der Begriffswörter, der Präfixe und Suffixe wurden später von anderen vermehrt, namentlich fanden nach dem Herrschendwerden der christlichen Religion viele Abkürzungen für biblische Namen Eingang. Während der Kaiserzeit war diese Schreibart in Rom sehr verbreitet und wurde selbst in Schulen gelehrt, doch sind nur wenige Texte davon erhalten. Mit dem Untergang der klassischen Sprache verlor sich auch das Verständnis der Noten, im Mittelalter kommen sie nur vereinzelt in notariellen Urkunden vor: die Kenntnis derselben wurde durch Wortverzeichnisse erhalten.