# Encoding of LATIN SMALL LETTER C WITH STROKE as a phonetic symbol

Peter Constable, Microsoft Corporation
2004-04-19

At the February 2004 meeting of the Unicode Technical Committee, a proposal was considered to encode the phonetic symbol LATIN SMALL LETTER C WITH STROKE. Some reservation was expressed on the part of some committee members, however, due to potential legacy encoding issues. A decision was made to give tentative approval of this character, but to prepare a public review issue to elicit feedback on the pros and cons of encoding this character. This document presents the pros and cons of encoding this character.

The question to which feedback is requested is this: *Do arguments for unification provide a strong enough case against arguments for encoding a new character that UTC should reverse it's decision to encode a new character?*

## Background

The text element *c-stroke* '¢' is often used in phonetic transcriptions to represent a voiceless alveolar affricate, particularly by Americanist linguists.

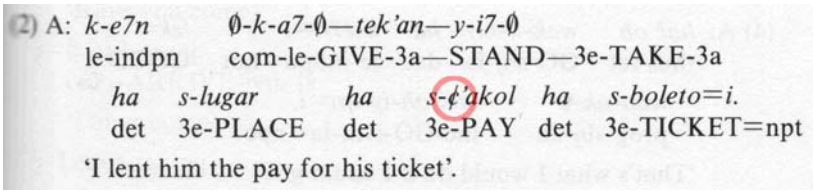
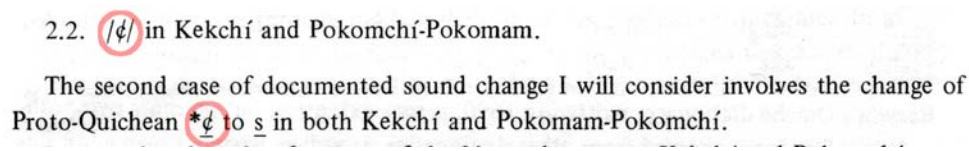
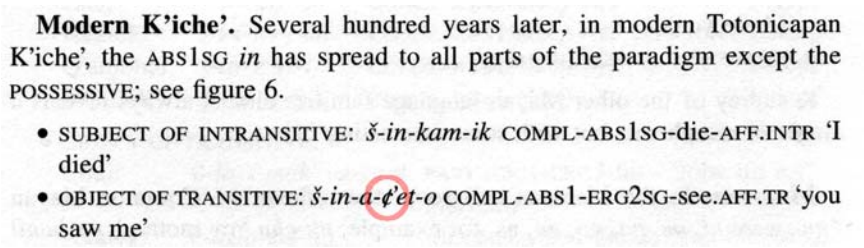

(2) A: *k-e7n* *∅-k-a7-∅—tek'an—y-i7-∅*
le-indpn com-le-GIVE-3a—STAND—3e-TAKE-3a
*ha* *s-lugar* *ha* *s-¢'akol* *ha* *s-boleto=i.*
det 3e-PLACE det 3e-PAY det 3e-TICKET=npt
'I lent him the pay for his ticket'.

Figure 1. From Brody (1986), p. 261.


2.2. /¢/ in Kekchí and Pokomchí-Pokomam.

The second case of documented sound change I will consider involves the change of Proto-Quichean *c̲ to s̲ in both Kekchí and Pokomam-Pokomchí.

Figure 2. From Campbell (1976), p. 124.



**Modern K'iche'.** Several hundred years later, in modern Totonicapan K'iche', the ABS1SG *in* has spread to all parts of the paradigm except the POSSESSIVE; see figure 6.

- SUBJECT OF INTRANSITIVE: *š-in-kam-ik* COMPL-ABS1SG-die-AFF.INTR 'I died'
- OBJECT OF TRANSITIVE: *š-in-a-ɋ'et-o* COMPL-ABS1-ERG2SG-see.AFF.TR 'you saw me'

Figure 3. From Robertson (1999), p. 457.

It is to be noted that this text element has similar appearance to the character U+00A2 CENT SIGN. The question, then, is whether to encode this text element as a distinct character, or to unify it with U+00A2.

## Arguments for and against encoding a new character

The primary arguments for encoding a distinct character are based on requirements for visual appearance and character properties.

The visual similarity mentioned above between c-stroke and U+00A2 CENT SIGN is to *one of the glyph variants* of CENT SIGN. That character has other glyph variants, however, such as "¢", "¢" and "¢", that are not acceptable for phonetic transcription.

More significantly, phonetic symbols often become adopted for orthographic uses in neo-literacy situations involving previously-unwritten languages. When phonetic symbols are adopted for orthographic use, innovation of an uppercase counterpart usually occurs as well. The character properties of U+00A2 CENT SIGN, however, are not appropriate for phonetic characters, given that potential for orthographic use. For instance, the CENT SIGN has a general category of *Sc* (currency symbol), which is not appropriate for a cased character.

These factors argue in favour of encoding a distinct character, and against unification with U+00A2. The argument *against* encoding a new character and for unification with U+00A2 CENT SIGN is based on potential concerns related to legacy data.

Up to the present, users have encoded documents that include the c-stroke text element. In those documents, this text element may well have been encoded as the character CENT SIGN. This would be the case if the following code points and the corresponding coded character sets (picking a few likely cases) were used:

| Code point | Coded Character Set |
|---|---|
| 0x9B | Code page 437, "IBM 'extended ASCII'" |
| 0xA2 | Code page 10000, "MacRoman" |
| 0xA2 | Code page 1252, "Windows 'ANSI' / Western" |
| U+00A2 | Unicode |

Table 1. Possible legacy encodings of the text element c-stroke

The argument against encoding a new character, then, is that it would create an interoperability issue between existing data that uses the coded character CENT SIGN and new data using a distinct coded character LATIN SMALL LETTER C WITH STROKE.

One possible rebuttal to that argument is that most existing data would have used a legacy coded character set (i.e. not Unicode), and that languages for which the c-stroke is used often also use other text elements that would not be supported in legacy industry character sets, such as the 'extended ASCII', MacRoman and Windows 'ANSI' code pages. Thus, most existing data would be encoded in terms of various non-standard character sets, and therefore any a priori identity between a particular coded character in those character sets and the character CENT SIGN cannot be assumed.

That rebuttal is, of course, of a theoretic nature. Far more relevant for users would be practical issues of data interchange. It is most likely that some existing data is defined in terms of non-standard encodings, in which case character conversion is already a necessity; but *also*, that some existing data is encoded in terms of an industry standard encoding using the character CENT SIGN, in which case encoding a new character introduces a need for data conversion that otherwise would not exist.

## Summary

Arguments for encoding the text element c-stroke as a new character and for unifying it with the existing character CENT SIGN have been presented. Users that work with writing systems that include the text element *c-stroke* are invited to provide feedback on these issues. *Does the argument for unification provide a strong case against encoding a new character? Are there other arguments for or against that should be considered?*

## References


Brody, Jill. 1986. "Repetition as a rhetorical and conversational device in Tojolobal (Mayan)." *International Journal of American Linguistics* 52.255-74.

Campbell, Lyle. 1977. *Quichean linguistic prehistory.* (*University of California publications in linguistics*, 81.) Berkeley, CA: University of California Press.

Robertson, John S. 1999. "The history of first-person singular in the Mayan languages." *International Journal of American Linguistics* 65.449–65.