From: Ken Lunde (lunde@adobe.com)
Date: December 16, 2015
Re: Hentaigana comment document

Nic,

The main concern that the UTC had with this proposal during the last meeting (early November) revolved around the characters that share the same source ideographs, have an "academic use" source, but no "family registration" source, and whose shapes are close from a calligraphic perspective. A good example of this is the pair JMJ-090108 and JMJ-090109, which have unique "academic use" sources, but only JMJ-090108 has a "family registration" source. Their strokes are fundamentally the same from a calligraphic perspective, and primarily differ in their vertical length, and as the vertical length increases, the more separate the strokes become. This is very similar to typeface design differences that can be observed in modern hiragana. Under this principle, I came up with 23 candidate pairs that may be suitable for unification. None of these 23 candidates have shared phonetic values, meaning that this is orthogonal to the suggestion that you made that resulted in a small number of unifications.

Would you be interested in taking a look at these 23 candidates? I think that your feedback on this particular issue would be helpful for all parties.

Another interesting case involves JMJ-090051 and JMJ-090052, which I would deem as unifiable, but because both characters have unique "family registration" sources, they would be encoded separately.

Also, I once again annotated the data file (attached) by highlighting in yellow or orange (orange is separate adjacent pairs or triplets) characters that share the same source ideograph, and further annotated the cases within each pair or triplet that have an "academic use" source but no "family registration" source by highlighting them in red.

I have added Debbie, the other Ken, and Kobayashi-san.

Regards...

-- Ken

From: Nicolas Tranter (n.tranter@sheffield.ac.uk)
Date: December 17, 2015
Re: Hentaigana comment document

Dear Ken,

With regards to the 2-page document that you sent, I should point out that etymologically a lot of the contested pairs and triplets also should include standard encoded hiragana. So, for example, JMJ-090117 (114), which I discuss later, is more like HIRAGANA TE て (U+3066) than like JMJ-090118 (115), and I'd argue that JMJ-090117 and JMJ-090118 are separate graphemes, whereas JMJ-090117 (114) and HIRAGANA TE are arguably not. At the end of this, I'll list all the queried proposed hentaigana with my brief thoughts on each. Before that, I'd like to raise the problems of the concept of "grapheme" and of relying on form and function as ways of deciding whether a particular item is a distinct hentaigana. Sorry for the length of this, but I'd like to make clear what the theoretical linguistic basis of my necessarily subjective opinions is. I'll use character names from the 20-page document rather than codes.

I hope this helps. Best wishes, Nic.

GRAPHEME:

The issue of what is a grapheme is complicated. This applies not just to the hentaigana but also to their source CJK character (or kanji, which I'll used hereunder). There are individual kanji that are officially recognised in Japan as separate characters and are encoded already in two forms, such as 萬 (U+842C) and 万 (U+4E07), or 禮 (U+79AE) and 礼 (U+793C). The simplified variant reflects long-term manuscript shorthand of the more complicated version. These two kanji pairs have different derived hentaigana associated with them: 萬 and MA-6, and 万 and MA-1, or 禮 and RE-1, and 礼 and RE-2 (although, a little inconsistently, the proposal gives 禮 as the source of both RE-1 and RE-2). Neither of these pairs of hentaigana are ones that are at issue because all four hentaigana are now Ministry of Justice name registration characters, but it illustrates the difficulties of using etymology as the basis for making distinctions.

Of the highlighted pairs and triplets of hentaigana, I would say that in a great many cases, although they have a common source kanji and a common phonetic value, they are separate graphemes because they tend to be used in the same manuscript and by the same hand. Although, especially in the earlier periods, you can see a spectrum of forms each subtly different from the next, certainly by the later periods even in relatively clear and consistent handwriting more than one form occurs, without any in-between transition forms. Two of the most obvious cases are KA-3 and KA-4 (both ka, both from U+53EF 可), or TE-5 or HIRAGANA TE て on the one hand and TE-6 (all te, all from U+3066 天). All these forms are frequent, often written in the same line of a manuscript by the same hand, and without the occurrence of transition forms.

I can illustrate this point in two ways.

Firstly, I attach a scan from an 1856 book that has jokes using the same repetitive formula: the word in the red oblong in each of the four occurrences is the same word (ka+ke+te), but in occurrences 1 and 2 the last kana has the form of HIRAGANA TE て and in occurrences 3 and 4 it is TE-6 - same reading (te), same source kanji (U+3066 天), written by the same hand, but kept quite distinct, and therefore separate graphemes as far as the writer was concerned. (Who knows if the writer even knew the etymology of all his hentaigana forms?)

Secondly, consider KA-3 and KA-4. Both have the same source, the kanji U+53EF 可. This character also constitutes the right-hand side of another separate character U+963F 阿, which is the source of the hentaigana A-3, yet this hentaigana's right-side is only identical to KA-3; it's never written with a right-side that looks like KA-4, which you would expect if KA-4 were used as a mere handwritten variant of KA-3. Again, I would argue that the writers of hentaigana clearly felt that KA-3 and KA-4 were separate graphemes.

As far as I'm aware, there has not been any serious attempt to study this approach to the definition of graphemes in Japanese. The lack of such a study - which would be an immense project covering 1,100 years of manuscripts - means that the decision on what variation to include and what not to include is very subjective.

The issue is more complicated because TE-5 and HIRAGANA TE are arguably not separate graphemes by the above definition, but just variants, albeit variants that occur in the same manuscript reflecting different degrees of care or speed of writing. If TE-5 was not already distinguished as a registration character it would be easy to argue that if TE-5 and HIRAGANA TE were to be distinguished, then forms with an extra slight squiggle in them (sorry for the lack of technical terminology here) should be distinguished too, such as one common variant of HIRAGANA SI or one common variant of HIRAGANA NO, neither of which are proposed here.

On the whole, I think NINJAL's decision on which hentaigana forms to include in the academic use list is

logical, although some forms are in linguistic terms "etic", i.e. not graphemic. Some are clearly in the same category as ka and te above.

FORM AND FUNCTION:

There is also a question over what constitutes a hentaigana and what is no more than a standardly hand-written version of its source kanji. Ideally, kana (hiragana, hentaigana) and kanji differ in both form and function: form, in that kana are more cursive and abbreviated than their source kanji; function, in that kanji are semantographic (or 'ideograms') and represent words or morphemes by meaning, while kana are phonographic and represent syllables with no indication of meaning. Unfortunately, (a.) kana and their source kanji, especially when the latter are relatively uncomplex, can be identical in handwritten premodern texts, such as HIRAGANA ME め and its source kanji U+5973 女, or hentaigana KA-3 and its source kanji U+53EF 可; (b.) medieval MSS are occasionally not averse to using a clearly kanji form with phonographic value; and (c.) kana are derived in any case from kanji forms, and the principle of using kanji phonographically was the device used in the Nara period before kana developed (so-called man'yōgana and senmyōgaki devices), and even now there are a small number of native words that can be written phonographically in kanji (a subset of so-called ateji).

In the proposed repertoire, there are forms that to me are no more than just how the character in question would be written in a manuscript. Keeping with the examples above, I'd cite TE-4, which I don't think I've encountered in late Edo period texts as a clear hentaigana, but which presumably occurs in earlier manuscripts. It is a good example of this, because in manuscripts the source kanji U+3066 天 written neatly would be identical to TE-4. It would be interesting to know what NINJAL's reasoning is on including TE-4.

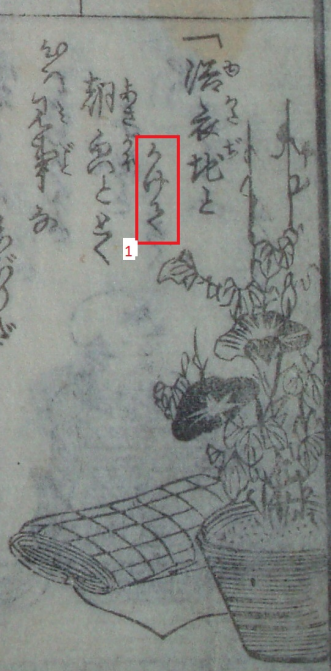
Another issue remains the Ministry of Justice's choice of standardized form for registration characters. In most cases, it's the MoJ's choices rather than NINJAL's choices that I would argue with.

So, my thoughts on each of the reddened characters. These are my very subjective opinions, and based on my familiarity with hentaigana primarily in the late Edo period. (I'm not that familiar with medieval manuscripts.) I write "emic" where I think the form in question is probably a separate grapheme in the same way as KA-3 and KA-4 are separate graphemes; "kanji used phonographically" in cases such as TE-4 above, where there may be a good reason for NINJAL to propose it but I personally do not see it as any different than a kanji in form; "Edo" if the form in question is the main hentaigana form or one of the main hentaigana forms in the late Edo period, and to replace it with a Ministry of Justice approved form would look strange. I write "etic" in those few cases where I'm not convinced of the need to distinguish the form. And I write "?" where I feel I can't begin to judge.

In my very subjective opinion, "etic" forms can easily be unified with either the MoJ's name registration form as you suggest or with a hiragana, and there's an argument for doing something similar with "kanji used phonographically". The "emic" and "Edo" forms should remain separate, I feel.

U-1: Edo
E-5: emic?
KA-4: emic
KI-2: ?kanji used phonographically
KU-1: kanji used phonographically
SA-3: Edo
SU-3: ?
SU-5: ?kanji used phonographically
SU-7: emic
SE-2: ?
SO-2: ? (more frequent in my experience as a handwritten premodern variant of U+6240 所 as a kanji; SO-2 is probably less common than SO-1 in this use)
SO-3: ?kanji used phonographically
SO-4: Edo
TU-2: ?etic (more squiggly form of HIRAGANA TU つ)
TE-4: kanji used phonographically
TE-6: emic
NA-5: Edo and emic
NA-8: ?emic (arguably NA-7 is a kanji used phonographically)
NE-3: ?
HA-5: emic
HA-8: ?
HI-7: Edo
HE-3: ?
HO-6: ?kanji used phonographically
MA-2: kanji used phonographically
MA-5: emic (formally it represents U+34BC 蔣, an abbreviation of U+6EFF 滿 with the latter's left-hand side

missing)
MI-3: kanji used phonographically
MI-5: ?
MO-2: ?etic (more squiggly form of HIRAGANA MO も)
MO-4: emic
YA-1, YA-2: ?
YA-5: etic
YU-3: maybe emic; certainly common in Edo texts
YO-4 ?emic (Edo texts frequently will contain two out of YO-4, YO-5 and HIRAGANA YO よ)
RA-4: probably not emic - a ligatured form of HIRAGANA RA ら
RI-2: Edo
RU-2: Edo
RU-4: Edo
RO-2: etic
WA-3: Edo
WI-2: Edo
WE-2: ?
WE-3: Edo
WO-1: ?emic
WO-2: ?
WO-7: Edo
N-MU-MO-1: ?kanji used phonographically


「浴衣地と
1


「月日と
2


「あの
3


「藝者と
4

From Ken Lunde (lunde@adobe.com)
Subject: Hentaigana/Hiragana unification candidates

Based on observations from Nicolas Tranter (see his 2015-12-17 email), I noted that the following hentaigana have the potential to be unified with their corresponding modern hiragana syllables, because the difference can be considered somewhat minor:

```
  1 HENTAIGANA LETTER A-1  = HIRAGANA LETTER A
 17 HENTAIGANA LETTER E-4  = HIRAGANA LETTER E
 24 HENTAIGANA LETTER KA-2 = HIRAGANA LETTER KA
 44 HENTAIGANA LETTER KU-2 = HIRAGANA LETTER KU
 63 HENTAIGANA LETTER SA-4 = HIRAGANA LETTER SA
 69 HENTAIGANA LETTER SI-2 = HIRAGANA LETTER SI
102 HENTAIGANA LETTER TI-5 = HIRAGANA LETTER TI
114 HENTAIGANA LETTER TE-5 = HIRAGANA LETTER TE
144 HENTAIGANA LETTER NU-2 = HIRAGANA LETTER NU
161 HENTAIGANA LETTER HA-4 = HIRAGANA LETTER HA
185 HENTAIGANA LETTER HE-7 = HIRAGANA LETTER HE
208 HENTAIGANA LETTER MU-1 = HIRAGANA LETTER MU
216 HENTAIGANA LETTER MO-2 = HIRAGANA LETTER MO
228 HENTAIGANA LETTER YU-2 = HIRAGANA LETTER YU
```

In particular, the representative glyph for HENTAIGANA LETTER SI-2 can be indistinguishable from HIRAGANA LETTER SI in some fonts, with Adobe's KazurakiSP2N-Light being one such example (Kazuraki on the left, hentaigana on the right):

That is all.