L2/13-075

Comments on the proposed Soyombo encoding

Shriramana Sharma, jamadagni-at-gmail-dot-com, India

2013-Apr-25

I have read through the earlier proposal for Soyombo L2/11-412 and the recent revised document L2/13-069 in detail. I have also had a direct telephonic discussion with the author Anshuman Pandey regarding the appropriate encoding model for the script.

In this document I wish to principally consider the encoding model proposed to represent consonant clusters. I also consider the SOYOMBO SYMBOL SVAYAMBHU.

§1. Consonant cluster representation

§1.1. Proposed Tibetan-like model

The proposal as it stands advocates encoding a full series of subjoined consonants. The rationale is that one of the intentions of the monk Zanabazar (Jñānavajra) in creating Soyombo was to transcribe Tibetan Buddhist texts, and since Tibetan is encoded using such a full subjoined consonant series, the same would be appropriate for Soyombo as well. This would help handle the frequent consonant clusters in Sanskrit and Tibetan without the need for a virama-type control character resulting in “decreased storage requirements” [TUS 6.2 on Tibetan, p 328 (358 of PDF)]. Further, cluster-initial RA LA SHA and SA take special forms in Soyombo, which might also easier be handled by separately encoding them as PREFIXED LETTER characters, which might then strengthen the case for encoding SUBJOINED LETTER characters for the cluster-non-initial consonants.

This model requires the encoding of 4 PREFIXED LETTERS and 40 SUBJOINED LETTERS.

§1.2. Problems with this model

One might consider the above model straightforward in representing the script as far as character-set to written-form mapping is concerned, which is true to an extent, but this model causes problems in various forms, which should be avoided.

§1.2.1. Encoding bloat

First is the **the encoding bloat in terms of number of codepoints used** for the additional characters not required by other models. One might say that “there’s lots of space in the SMP”, but this space is fast filling up with (and is needed for) old scripts that ongoing research (including my own) uncovers and encodes. It is inappropriate to overpopulate this space with characters which can be avoided.

§1.2.2. Unnecessary complications in input methods

Second, such a model does not help the development of simple input methods for the script. Given that these encodings of historic scripts are in the end principally intended for the storage of old documents composed in them, and someone has to input each individual character manually, one must try to make things convenient for that.

As there are a limited number of keys in a keyboard, it is certain that only the basic set of consonant letters can be directly mapped to the keys (even permitting the use of Shift). Requiring use of additional modifier keys like Ctrl/Alt/Shift/Meta combinations would be too complicated for usability in the case of such frequent use characters.

Due to this, what existing Tibetan input methods on popular platforms (source: http://writecantonese8.wordpress.com for MS Windows 7, Vinodh Rajan for Mac OS X, I wrote one for Win XP myself) do is to emulate the input of virama-based encodings and have a particular chosen key input a dummy keycode which would cause the next consonant keystroke to input the SUBJOINED LETTERS. This produces the unnecessary complication of mapping <dummy_keycode> + <keystroke_of_regular_consonant> to the SUBJOINED LETTERS. One could have gone with a virama-based encoding in the first place and have the smartfont take care of the rest. (Remember that the smartfont must do the appropriate substitution/positioning even in the case of SUBJOINED LETTERS.)

§1.3. Alternative Model 1 – CONSONANT JOINER model

The obvious alternative model to encoding SUBJOINED LETTERS is to encode a virama-like character which would cause the appropriate clustering behaviour just like in South Asian and Southeast Asian scripts.

One notes that there is a difference between the South Asian and Southeast Asian approach to such a character:

In South Asian scripts, a visible virama is frequently used and clusters are only of consonants. So a single character is used for the visible virama and also has the additional function of joining consonants. In Southeast Asian scripts, a visible virama is rarely used, and the clustering behaviour is (IIUC) not limited to consonants, so the visible vowel killer and the subjoining functions are separated into two different characters.

As far as Soyombo is concerned, no visible virama is used. Even (the valid) syllable final consonants are denoted by distinct marks astride the bottom of the frame stem. Therefore there is no scope for a character representing a visible virama. However, one can encode an invisible control character Soyombo Consonant Joiner which will only take a visible fallback form such as a dotted square when not followed by an appropriate letter.

A consonant C2 following this Joiner as C1 + Joiner + C2 will be rendered as a subjoined glyph just as in South Asian and Southeast Asian scripts. When however, C1 is one of RA LA SHA SA, it will take the attested prefix form and C2 will be displayed in the nominal form, similar to the reph-formation of South Asian scripts. Thus one can avoid the disadvantages caused by encoding the 4 prefixed and 40 subjoined consonants.

In Tibetan, there is attested contrast between full and reduced form of some of the "semivowel" consonants YA RA LA VA in superfixed and subjoined positions. With a view to maximizing the capability of one-to-one transliteration from Tibetan, in Mongolian Square such reduced forms are proposed to be encoded separately, with the Mongolian Square Subjoiner always producing only subjoined *full* forms of the consonants (L2/13-068). However, in Soyombo, even though it was also designed to be able to write Tibetan, the situation is different. There is only one frame per syllable and the consonant nuclei are appropriately shaped and positioned within it. A consonant at a given position in the cluster is always represented by the same written form and there is no contrast such as is seen in Tibetan. A Soyombo Joiner can hence well take care of all consonant clusters.

Though this model gives up the advantage of "decreased storage requirements" cited for (Tibetan and) the Subjoiner Letters model, in this age of cheap storage this is a non-issue, especially seeing as encoding bloat is avoided, and implementation is simplified.

While it is true that one has to rely on smart fonts to render the C1 + Joiner + C2 sequences properly, one notes that such smart behaviour (as in positioning, resizing etc) is needed for Soyombo even in the Subjoined Letter model. The presence of syllable final consonant signs and the shaping behaviour noted in §5 of the proposal also requires this.

§1.4. Alternative Model 2 – frame-nucleus model

The principal opposition to the Joiner model (as voiced by the proposal author), that introducing such an invisible Joiner is foreign to the native user's perception of the script, especially seeing as there is no written virama in the script. Further, the native user perceives the frame as an integral part of a written letter, and having an invisible character remove it is somehow jarring to that perception. This is why the proposal states in advocating the Subjoined Letters model:

> *"… the subjoined-letter model complements the frame-nucleus structure of the script and adheres to the method of writing Soyombo by hand."*

Apparently it is felt better to have distinct frameless nuclei encoded as separate characters and attach them to a frame-nucleus combo, rather than have the frame magically removed!

Such subjective issues have been encountered in some Southeast Asian scripts as well, where the encoding of such a special joiner was strongly opposed. However, the encoding exists to best serve the needs of the natives even if it involves some effort of comprehension on the part of the natives that writing by hand and encoding as digital text are not exactly the same thing.

So there isn't really anything objectionable in the Joiner model. Despite this, the glyphic structure of Soyombo inspires one to think of yet another model which, though probably unique to Soyombo, would be very intuitive and straightforward to implement.

We have noted that all Soyombo syllables have a frame consisting of a down-pointing triangle on top and a stem on the right (samples on next page). Within this is placed the nucleus, "a distinctive element that represents a phonetic value". Multiple nuclei representing consonants in a cluster are all stacked within the frame. Some of these, RA LA SHA SA to be precise, change their form when cluster-initial. Vowel signs attach to various parts of the frame, and final consonant signs are attached to the bottom of the frame stem. (This is just a restatement of §4.2 from the proposal.)

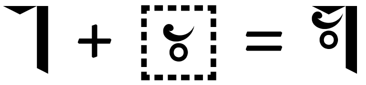
Thus the frame has a structural value and the nuclei and the various signs have phonetic values. Perhaps one might even consider the frame to represent the inherent vowel, whereby the nuclei would represent pure consonants.

The current proposal has suggested that each frame-nucleus combo representing a basic letter with inherent vowel be encoded as a single character (and this is retained in the Joiner model) and that the basic consonant nuclei alone may then be separately encoded as the Subjoined Letters to be used in clusters (which is avoided in the Joiner model).

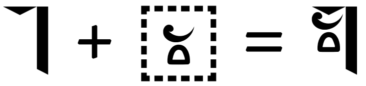
Now the distinction of the regular Letters vs the proposed Subjoined Letters is essentially only the absence of the frame in the latter. It then would follow that if one encodes the frame alone as Mongolian Square Frame (with GC=Lo), one can merely encode the nuclei for all the consonants once and leave it at that.

The syllable encoding model will now closely mirror the actual writing:

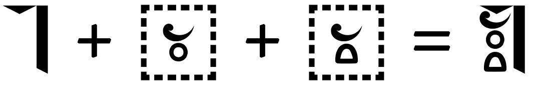
Frame + C1-Nucleus [+ C2-Nucleus ...] [+ Combining Marks]



Frame + MA-Nucleus = MA



Frame + YA-Nucleus = YA



Frame + MA-Nucleus + YA-Nucleus = M·YA

This is even more straightforward than the Subjoined Letters model, because you are merely asking the user to encode what s/he writes: a frame first, then as many nuclei as required for the cluster being represented, then the requisite vowel signs etc.

It is of course up to the smart font to properly position the nuclei, remove the swash from any subjoining YA or RA nucleus, and modify the frame stem as required (to match height of stack etc, see §5 of the proposal). In the same process, the smart font programming can also identify the RA LA SHA SA nuclei at cluster initial position, and substitute them by their reduced prefixed forms; so this model avoids the need for separate Prefixed Letter characters just like the Joiner model.

However, in this model, obviously no Joiner is required to remove the frame for the non-initial consonants, because a cluster will have only one distinctly input frame at the start. Thus this model both avoids introducing an unfamiliar invisible control character and also adheres to the native user's perception of the writing process. As such, it may well found to be the most appropriate.

Shagdarsürüng, one of the chief references for the proposal, has even illustrated the broken-up components of writing Soyombo (p 37), which could lend support for this model.

We have to consider one final aspect of this model, though, and that is regarding the storage requirements. While we have already rejected this as a non-issue, and nothing has changed since then, we still consider it if only for academic completeness.

The frame-nucleus model is more economical than the JOINER model in representing consonant clusters, since in the place of multiple JOINER characters, one only has to input an single FRAME character at the start of the syllable. (Of course, for two-consonant clusters, one FRAME vs one JOINER doesn't make a difference.)

However in the case of syllables which do not involve consonant clusters, the frame-nucleus model is costlier than the JOINER model by one character-space per syllable, because the frame has to be input separately. Thus the total equation depends on the ratio of syllables with consonant clusters to those without. Considering that a significant portion of the usage of Soyombo is for Sanskrit and Tibetan texts which would have a significant percentage of syllables *with* clusters, I think it is safe to say that the overall storage requirement of the frame-nucleus model will not be too much in excess. And given the circumstances this may be a small price to pay to conform to native users' perception.

(The SUBJOINED LETTERS model is of course the most economical for storage, but it sacrifices economy in encoding instead.)

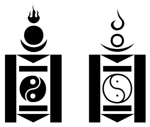
§1.5. Summary of consonant cluster model

Encoding a full set of SUBJOINED LETTERS and some PREFIXED LETTERS is found superfluous, bloated and problematic. It is recommended to replace this model with one of:

1) the JOINER model, where the regular consonants (as frame-nucleus-combo) are encoded, and which uses an invisible virama-like specialized JOINER which will join the preceding and succeeding consonants in the cluster; the disadvantage here is being dissatisfying to native perception of the script;
2) or the frame-nucleus model where a distinct FRAME plus 40 consonant nuclei are encoded, and appropriate subjoined or prefixed forms are produced by smart fonts handling the sequence of nuclei; the disadvantage here probably being a somewhat increase in storage requirements.

Native users/experts should decide. If they can accept an invisible JOINER, then in keeping with the majority of Brahmi-derived scripts, that model may be chosen. If one would prefer to avoid an invisible JOINER, then the frame-nucleus model can be chosen.

§2. Soyombo Symbol Svayambhu



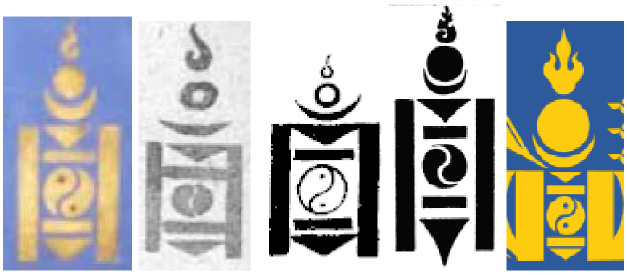
The two above glyphs are proposed to be encoded as two distinct characters, the first as the Soyombo Symbol Svayambhu and the second as the Soyombo Head Mark. However, they appear to be stylistic variants of the same character originally design by Zanabazar.

The proposal states (in §4.13) that the difference between these two is in the top, i.e. the shape of the flame. The form officially recognized as the National Symbol of Mongolia apparently has a three-tongued flame. The other form has a single tongue flame.

Other difference between the two forms include: the bindu part of the candrabindu and the taijitu* being filled vs unfilled, the size of the bindu and the orientation of the taijitu: rising on the left vs right. Apparently** while the taijitu is most commonly presented as rising on the left, other forms are also used (as in the flag of South Korea). Therefore these are merely variants which do not warrant distinct encoding.

Further, the sources listed in the proposal: "Histoire du livre" (p 19), Kapaj (p 20) and Shagdarsürüng (p 22, 23) all show only a single symbol named Soyombo listed before the first letter of the alphabet, and the crown-like terminal mark at the end. No distinct head mark apart from the Soyombo symbol is listed.

Further, these symbols all show variations especially w.r.t. the shape of the flame, the filled/unfilled parts and the proportions. Given this, the distinction pointed out by the author regarding the shape of the flame is irrelevant.



* yin-yang symbol, see: http://en.wikipedia.org/wiki/Taijitu. Zanabazar appears to have created the Svayambhu symbol out of a fusing of an OM (hence the candrabindu) and the taijitu.

** Source: same Wikipedia article linked above, retrieved 2013-Apr-25

It would seem that the distinction has been taken by the author from the Soyombo font that he used. The designer of this font notes that "a variant of the opening symbol of the Soyombo script ... became the national symbol of Mongolia". (See http://userpage.fu-berlin.de/corff/im/Soyombo/overview.Soyombo.html) and has provided glyphs of both styles at separate legacy codepoints.

However, since none of the sources attest a separate head mark, and even in the Soyombo symbol many variations are seen, one cannot justify encode a separate character for the head mark apart from the symbol.

It is understood that the official form of the symbol bears a certain prescribed distinctive shape which may not be found in manuscripts. (Indeed, none of the sources attest the exact symbol used.) Nevertheless, this need not be disunified. Such a formalized form may be represented as this same character by an appropriate font.

I also suggest that keeping in with the Mongolian tradition, the Soyombo symbol be moved to the head of the proposed block.

-o-o-o-