JTC1/SC2/WG2 N3637

TITLE: Outstanding Issues on Old Hungarian/Szekler-Hungarian Rovas/Hungarian Native Writing
SOURCE: Deborah Anderson, SEI, UC Berkeley
STATUS: Expert contribution
ACTION: For consideration by WG2
DISTRIBUTION: ISO/IEC JTC1/SC2/WG2
DATE: 22 April 2009

The following is a summary of the outstanding issues on the various Hungarian Runic proposals, based on documents by Gábor Hosszú ["GH"], Gábor Bakonyi ["GB"], and Everson/Szelp ["ME/SS"].

Source Documents (Proposals):

N3527 Szekler-Hungarian Rovas by Gábor Hosszú 2008-10-04
N3566R Hungarian Native Writing by Gábor Bakonyi 2009-02-05
N3615 Revised proposal for encoding the Old Hungarian script by Michael Everson and André Szabolcs Szelp 2009-04-16

Other docs submitted to WG2:

N3532 Mapping between two Old Hungarian proposals by Michael Everson and André Szabolcs Szelp (Note: This is based on the earlier version of N3615, so it is not up-to-date)

1. Name of the script

ME/SS: Old Hungarian
GH: Szekler-Hungarian Rovas
GB: Hungarian Native Writing

2. Location of script

Note: The BMP is almost completely full, so historic and modern scripts are now being allocated to other planes. It is best to have scripts located in one single block, and not broken up across blocks. When selecting an area on the Roadmap for a script to be encoded, the block must be beside other scripts that share the same directionality. In other words, a RTL script must be located in an area of the Roadmap where other RTL scripts are located on either side.

ME: SMP (10C80-10CFF)
GH: Partly on BMP and partly on SMP (BMP: letters 0860-087F, 1C80-1CCF, AB70-AB7F, AB80-AB8F, numbers and punctuation 1AB0-1ABF; SMP: historical ligatures 10C60-10CDF)
GB: BMP (0860-0897)

3. Case distinction

Note: Evidence for casing is provided on p. figure 11-5 (GH) and figure 11 (ME/SS). Uppercase letters appear to be used today under influence of the Lain script, and can be used if desired.

Upper and lowercase characters are contained in the proposals by ME/SS and GH, but not GB.

4. Character Repertoire:

a. Bug characters

There are a number of disagreements on the "bug" characters (are they ligatures and which are variants of others, if any?). As these are historic characters, it is not always easy to determine details on their background and exact values (or names). Can evidence be provided showing the variants of the same character appearing separately on the same page and hence needing to be separately encoded?

Views:

GH: encode 12 "bug" characters, and does not consider them variants of one another or ligatures:

ME: encodes 7 of them

GB: Just encode 6 most common

7 Agreed Upon Characters (ME/SS and GH):

amb , and  , emp  , ent  , tprus (/ent-shaped sign)  , unk  , us 

Disagreement on Characters

ME/SS Consider as Variants
ant  (variant of ent )
nb  mb  tpru  (variants of emp )

ME/SS Consider Ligature
nap  (consider ligature of N - P)

b. Ligatures (not including bug characters)

Note: For the Unicode policy on ligatures, see http://www.unicode.org/faq/ligature_digraph.html. Ligatures seem to be productive in this script. The approach taken should offer a means to create the combinations that users need. This involves ensuring the base characters are encoded, and documenting the sets of letters that ligate so fonts can properly create the forms.

ME/SS: don't encode, instead handle ligatures using OpenType features or U+200D Zero-Width Joiner to join the two letters; examples of 72 ligatures with ZWJ are listed in N3615 on pp. 7-8 of N3615.
Besides various letter combinations, four single letter ligatures are noted by ME/SS: Q (ligature of K-V), W (ligature of V-V), X (ligature of K-Sz), and Y (ligature of J-I).

Example of ligature from proposal of ME/SS:

GB**:** use of Zero Width Joiner to create ligatures
GH: proposes encoding 64 small and 64 capital historical ligatures as separate characters.

c. Numbers

Note: For encoding, it is important to show evidence of a given character is use in printed materials, this is especially true for very new additions.

ME/SS, GH, GB all include characters for: 1, 5, 10, 50, 100, 1000), but GH has an additional character, 500.
Note on '500' from p. 16 of GH: Scholar Atilla Koricsánszky had stated '500' had to exist, so at a meeting in 2008 a glyph was created by Tamás Rumi.

d. Punctuation marks

Note: In general, punctuation marks may be used across different scripts. Currently encoded characters are in ALL CAPS in the table below with the codepoints.

The table below lists possible characters for use in this script based on a rough description of their shape or use.

PROPOSED CHARACTERS WITH AGREEMENT BETWEEN AT LEAST TWO PROPOSALS

ME/SS	GH	GB
propose reversed comma	proposes reversed comma	proposes reversed comma
propose double reversed low quotation marks (for beginning quotations)	proposes double reversed low quotation marks (for beginning quotations)	

USE ALREADY ENCODED CHARACTERS?

ME/SS	GH	GB
0020 SPACE		0020 SPACE
2E31 WORD SEPARATOR MIDDLE DOT	proposes single dot different from 2E31 due to variants	proposes single dot
205A TWO DOT PUNCTUATION		proposes double dot
205D TRICOLON		proposes three dots (which he suggests be used for question mark)
205E VERTICAL FOUR DOTS		proposes four dots
0021 EXCLAMATION MARK		
002D HYPHEN-MINUS		
002E FULL STOP		
003A COLON		
201F DOUBLE HIGH-REVERSED-9 QUOTATION MARK	proposes high ending double quotation mark (for ending quotations)	
204F REVERSED SEMICOLON	proposes reversed semicolon	
2E2E REVERSED QUESTION MARK	proposes question mark	
002F SOLIDUS		proposes / (used as colon)
0304 COMBINING MACRON		proposes combining macron (used as long-mark)
0307 COMBINING DOT ABOVE		proposes combining dot (used as short-mark)
2E17 OBLIQUE HYPHEN		propose double hyphen (used to break a word)

e. Other Characters (not bugs or ligatures)

i. Letter A

ME/SS: Letter A changed (N3615) to:

BOLOGNA A, NIKOLSBURG A, CSULYAK A

	Adorján Magyar	Győző Libisch	Sándor Forrai
bol$_A$ ~	*a* /ɒ/ ~	*a* /ɒ/ ~	*a* /ɒ/ ~
nik$_A$ ~		*á* /a:/ ~	*á* /a:/ ~
csu$_A$ ~	*á* /a:/ ~		

GH: A  AA 

GB: A 

ii. Glyph variants? Separate characters?

ME/SS	GH	GH separately encodes:
ES 	S 	AS  (but considered glyph variant by ME/SS)
ETY 	TY 	ATY  (but considered glyph variant by ME/SS)
 *NIKOLSBURG UE*	*UEE*	*OLD OE 2*  (based on N3532)
 *RUDIMENTA UE*	*UE*	*OLD UEE*  (based on N3532)
 *OEE*	*OEE*	*OLD OEE*  (based on N3532)

5. Character Naming Issues (focusing on ME/SS and GH)

ME/SS	GH	GB
Precede consonant name with vowel: EB, EC, EH	Ok to start name with consonant: B, C, H (24 letters)	
 NIKOLSBURG OE	OLD OE	
 RUDIMENTA OE	OE	
NIKOLSBURG ETY	HH	
 NIKOLSBURG UE	UEE	
 RUDIMENTA UE	UE	
 "ENT-SHAPED SIGN"	"TPRUS"	"TPRUS"

6. Sorting Order

Note: The order of characters that appears in the charts and names list does not define sorting order. Sorting order can be tailored.

In general, the Latin ABC order is followed in the three proposals, with long vowels following the short vowels. The main difference between the versions is where to put the historic characters and the homorganic nasals (AMB, ENC, AND, UNK, EMP, ENT).

- Should the historic letters and homorganic nasals be interfiled with the other letters or should they come together after the main set?
- Are there handbooks or alphabets with the historic characters interfiled?
  If not, can any guidelines be put forward?