Doc Type: Working Group Document
Title: Proposal to encode Latin letters used in the Former Soviet Union
Authors: Nurlan Joomagueldinov, Karl Pentzlin, Ilya Yevlampiev
Status: Expert Contribution
Action: For consideration by JTC1/SC2/WG2 and UTC
Date: 2011-10-18

Terminology used in this document:

"Descender" refers to the specially formed appendage on letters like the one in the already encoded letter U+A790 LATIN CAPITAL LETTER N WITH DESCENDER.

"Typographical descender" refers to the part of a letter below the baseline, thus resembling the term "descender" as used in typography.

1. Introduction

In the wake of the October Revolution of 1917 in Russia, alphabetization of the people living in the then formed Soviet Union became an important point of the political agenda. At that time, some languages spoken in the Soviet Union had no standardized orthography at all, while others (especially in areas where the Islam was the predominant religion) used the Arabic script. As most of these orthographies did not reflect the phonetics of these languages very well, and as the Arabic script was considered unnecessarily difficult by some due to its structure, for most of the non-Slavic languages it was decided to design new orthographies from scratch. An important event was the Turkologic congress in Baku of 1926 where the Jaꞑalif alphabet was passed. Following this decision, Latin orthographies for several other languages were designed in the following years.

The Latin orthographies were in use as the primary orthographies of these languages, and several books, magazines, and newspapers were printed using them.

They were abandoned from 1938 on, as Stalin requested the use of the Cyrillic script for all these languages after a decision that all citizens of the Soviet Union have to learn Russian. The Cyrillic orthographies designed then are mostly the ones still in use today, except where they are replaced again by Latin orthographies in the countries which do no longer belong to Russia after the decease of the Soviet Union.

Detailed information is found in [4] and [5] (see section 5. "References").

An early summary of the Latin letters designed for the Latin orthographies which are subject of this proposals is found in [1] (Yushmanov, see section 5. "References"). This proposal highly relies on this compilation and presents the characters with references to it, see fig. 1 to fig. 3. (However, for all characters proposed here we show actual use; no character is proposed based only on its listing in Yushmanov. Also, the languages listed for each character are based on the evidence presented here, rather than citing Yushmanov's listings.)

The characters proposed here are primarily needed by librarians and historians, to enable the correct referencing and citing of the numerous original works using the Latin orthographies.
2. Encoding Considerations

All characters proposed here are Latin letters without special properties, mostly in case pairs. A few are lowercase only, as two orthographies (Adyghe and Abkhaz) are monocameral.

2.1 The descender

A common feature of the letters newly created for the 1020s/1930s Latin orthographies of the Soviet Union is the descender. It is an appendage put on the bottom of the letter to create a new one, which usually is phonetically related to the original one. Obviously, it has its model in the appendage which distinguishes the Cyrillic letters Ш (SHA) and Щ (SHCHA).

In the Latin orthographies, it has no precedence and therefore no typographic tradition. Thus, in the evidence presented here it shows many different forms (rectangle, triangle, Щ-descender, hook similar to a palatal hook, cedilla, or comma below). Most often, it is placed at the right end of the bottom of the base character, but small right displacement from the central line or exact central placements also occur (especially on letters with round bottom like C/D/S). Usually (but not in every case), it is more prominent than the Ш/Щ difference in common Cyrillic fonts, probably to support eyes which are not accommodated to the Ш/Щ difference.

This variation occurs within samples of the same orthography and even also within the same book (see e.g. fig. Komi-3).

Therefore, all forms of the same Latin base letter with descender are unified.

In accordance with the already encoded Latin H/K/N/Z with descender, and as the descender has no characteristic shape by itself, all letters with descender are proposed as individual atomic letters, rather than to propose a new diacritical mark "DESCENDER BELOW".

This is also done for C and S. These letters, while occasionally show a form of the descender resembling a cedilla or comma below, in most of the specimens show a form resembling the descender form used for the other letters (e.g. triangle). Also, the horizontal placement is right from the vertical center in many cases (it is even at the extreme right in the Adyghe samples). The Komi examples show such variations even in the same book.

See also fig. Kabardian-3.

2.2 Glyph variations in vertical extension

A feature which distinguishes the Latin script from the Cyrillic script is the fact that Latin letters which end in a vertical line or an open bow at the bottom, are commonly extended to the descender line in Latin (i.e. they have a typographical descender), while they do not this consequently in the Cyrillic script (i.e. they show the shape of the capital letter, only being reduced to x-height).

Thus, when Latin letters are modeled after the Cyrillic case pairs Ъ/ѣ, they will look Ъ/ѣ (showing "long" glyphs for the lowercase letters; consequently done also in another area as shown in fig. 7 and 8).

In fact, such "long" glyphs are found, while the "short" glyphs are also found. The same is the case for similar letters like the "mirrored ghe"s Ꞁ/Ѿ Ꞁ/Ѽ, the already encoded Ѥ/Ԉ, or the already encoded Ƣ/ƣ for which glyphs like Ƣ/ƣ occur.

For these letters, the "long" or "true Latin" forms (which show the typographic descenders) are proposed for the representative glyphs, while the "short" forms are considered as glyph variants.
In no case, we have proposed different letters only to separate the "long" from the "short" forms. Consequently, we have unified the "short" Ɂ with the already encoded "ѣ" (U+0292).
In the same way, we have unified the "short h with hook", which has no real ascender (see fig. Kabardian-2) with the already encoded "и".

2.3 Other glyph variations within the 1920s/1930s Latin orthographies

As the users of the newly orthographies were not used to the typographical traditions of the Latin script, while some were familiar with the Cyrillic script, some glyph variants occurred which may look strange for people familiar with the Latin script.

The capital Y often shows a form resembling the Cyrillic У. This form obviously matches the common form of the lowercase y.

The capital D occasionally shows a form following the curvature direction of the lowercase d, thus appearing mirrored to users familiar with the Latin tradition (see fig. Udi-1).

In the area where Komi is spoken, such glyph variants also show a little ascender, obviously influenced by the Cyrillic letters U+0500/0501 Ԁ/ԁ which were in use 1918-1930 predating the introduction of the Latin alphabet in that area.

Such glyph variants are treated as such and are not proposed here as new letters.

2.4 Character forms similar to Cyrillic ones

The Latin and Cyrillic alphabets share several letters which commonly have identical forms, independent whether they usually denote the same phoneme (A/a) or different ones (P/p). Also, in several cases, the complete case pair resembles (e.g. C/c), while in other cases only the capital letters resemble (e.g. T/t vs. T/T).

Thus, it is no surprise that some of the additional letters proposed here also resemble Cyrillic letterforms.

However, for complete case pairs, this is only true for the following ones: Ё/ё, Ь/ь, Х/х, Х/х.

(Additionally, the pair Ё/ё vs. Ѕ/ɨ could be called resembling by glyph variation.)

This leads to the possibility of unifying the Latin characters proposed here with the existing Cyrillic letters, to save a few code points.

However, we consider such script mixing being a bad thing for the following reasons:

- The issue of identically looking letters in different scripts is understood and handled very well already.
- Script mixing within one orthography (i.e. within one locale) makes issues like sorting and searching (and specifying the rules for this) unnecessary complicated.
- Script mixing within one orthography hampers spoofing protection in IDNs (like URLs or e-mail addresses) based on disallowing script mixing (like accessing an URL "google.com" where the second letter is a Greek omicron and the third one is a Cyrillic small o). The fact that users will create URLs containing letters of the orthographies considered here is not confined to the theoretical possibility that such orthographies will be revived. It is expected that historians or individuals will create archive pages using the original titles of publications using these orthographies, e.g. (see the newspaper title in fig. Kazhak-1) www.mydiscussionpage.xy/kazhak_history/newspapapers/sotsijalds_qazaqstan.html.
• Script mixing irritates the common user, as when they know that they have to use Cyrillic letters in Latin text anyway, they will do so when the outcome shows correct to them. Thus, they will use Cyrillic character even when an appropriate Latin one exists (e.g. Cyrillic small ь instead of Latin small capital b, to give an example not as trivial as mixing Cyrillic П/п with Latin P/p in the same text although in fact only the latter one was yielded).

• Regarding Х/х which are independent straightforward derivations of the respective Latin and Cyrillic base letters, it would be especially irritating if the derived letters were unified across scripts, while the base letters are not.

• Besides the heritage of Unicode 1.0, script mixing within European Alphabetic Scripts was done in no case for orthographies (i.e. character sets which are intended for use in everyday life, as opposed to scientific sets like the Uralic Phonetic Alphabet). As a recent example, U+A698/A699 CYRILLIC CAPITAL/SMALL LETTER DOUBLE О were accepted (see PDAM 1.2, WG2 N4107), instead of unifying them with U+A74E/A74F LATIN CAPITAL/SMALL LETTER О which exactly look the same.

2.4.1 The special case of Latin Ь/ь

There exists a letter pair U+0184/0185 LATIN CAPITAL/SMALL LETTER TONE SIX, which is used for a historic orthography of Zhuang (used 1957-1982) which was claimed in earlier discussions on the Unicode mailing list to be identical to the Latin Ь/ь.

This issue can be resolved only by thorough examination and discussion of the glyph variation which this letter has in Zhuang source material. This is beyond of the scope and the resources of the work presented here, and therefore will be discussed in a separate paper at a later time.

The possible outcomes of such a discussion are:

• Either, the Latin Ь/ь will be unified with the Zhuang tone six, and the representative glyph will be changed resembling the Cyrillic Ь/ь, as such glyphs are used in several orthographies (Jaalif and most of the orthographies considered here), while Zhuang then is just one orthography among many using this character.

• Or, such a glyph change is not appropriate for Zhuang. This means that we have different letters here, as the glyph variations do not overlap considerably, and the Latin Ь/ь gets encoded as a new true Latin letter.

2.5 Casing issues

Special casing (tailoring; like for Turkish İ/İ – İ/i) is to be used anyway for most of the orthographies considered here, as they pair U+0042 LATIN LETTER CAPITAL B with U+0299 LATIN LETTER SMALL CAPITAL B as lowercase form. This was decided for the Jaalif alphabet which acted as precedent for several of these orthographies.

Thus, it imposes no additional complexity if existing forms are paired with newly encoded letters, thus avoiding the encoding of letters looking identical to existing ones only to get complete case pairs which are marked as such unambiguously by Unicode properties.

2.5.1 Capital and lowercase forms resembles existing letters, but are not paired

In this case, no new letters are proposed, as such cases can be handled by tailoring.
2.5.2 The capital form resembles an existing letter, while the lowercase form does not

In this case, only the lowercase letter is proposed as a new character, and the casing has to be handled by tailoring.

In this proposal, this applies to the case pairs:
U+A780 LATIN CAPITAL LETTER TURNED L is the uppercase form for the proposed LATIN SMALL LETTER REVERSED GHE.
U+A7FB LATIN EPIGRAPHIC LETTER REVERSED F, although having the property Lo rather than Lu, is the uppercase form of the proposed LATIN SMALL LETTER REVERSED F.

2.5.3 The lowercase form resembles an existing letter, while the capital form does not

If the existing lowercase letter has no capital counterpart until now, this imposes no problem as the newly encoded capital letter is simply to be formally paired with the existing lowercase letter. (However, this proposal does not contain such capital letters.)

Otherwise (if the existing lowercase letter already has a capital counterpart, which does not resemble the capital letter proposed here), only the capital letter is proposed as a new character, and the casing has to be handled by tailoring.

However, having capital letters (declared by property Lu) without lowercase counterparts imposes problems, as due to stability policies it is not possible to declare lowercase counterparts when they are finally included in a Unicode version. Also, they could not be used in any IDNs (e.g. URLs or e-mail addresses), as they had to be mapped on their non-existing lowercase counterparts there.

To circumvent any such issues, they are treated as follows:

- They are mapped to their lowercase counterpart formally. This has the consequence that casing is not a roundtrip operation for this letter, as lowercasing and then uppercasing yields the original capital counterpart of the associated lowercase letter. This follows the precedent of U+1E9E LATIN CAPITAL LETTER SHARP S.

- They are given a compatibility equivalence to the original capital counterpart of the associated lowercase letter. This circumvents all problems with IDNs (URLs and e-mail addresses), as the newly encoded letter and the original capital counterpart of the associated lowercase letter are thus declared to be equivalent when used there.

- They are placed into a specific group, by proposing to be placed in the “Alphabetic Presentation Forms” block. As true presentation forms (like U+FB16 LATIN SMALL LIGATURE ST) will no longer be encoded, the gaps in this block are considered to be now the appropriate place for letters which are special by having a compatibility equivalence to other letters.

It is noted that, in spite of the compatibility equivalences given for technical reasons, the capital forms are full-fledged letters by their own, different from the already encoded letters which have developed in unrelated areas and eras.

They are not glyph variants. Glyph variation is something which must be recognized by the users; e.g. the $g/g$ variation (or the variation between triangle-formed and hook-like descenders) appears to the same group of users. But the groups of common users of Iľ and ăľ are divided by geography and time, and they never see the other group’s character in ordinary life. Even if they could decipher it correctly, they would consider it wrong, like an English speaker would look at an "Ə" found in the place of an "E".

Proposal to encode Latin letters used in the Former Soviet Union
2011-10-18
Being forced to use African letters like Ɥ/ Ɥ/ Ɥ/ Ɥ instead of Ꞇ/ ꞇ/ ꞈ/ ꞉ means having to use a cipher rather than an encoding.

The fact that such ciphering is supported by font technology (like it was used to write Greek in the 8-bit area by using special fonts which mapped Greek letters on ASCII code points), does not change this. It is dangerous to burden the users e.g. of bibliographical databases to force to care about font selection only to get their database records presented correctly (as they could get wrong presentations when using fonts with the wrong glyphs but being correct in Unicode terms).

2.6 Special naming issues

The capital forms of "h" and related letters (turned h, heng) which are based on the common lowercase form are named "wide", as they usually are distinguished by their considerably wide appearance from the lowercase ones (see e.g. fig. Abaza-6 for the WIDE H WITH LEFT HOOK, or fig. Ingush-1 for the WIDE HENG).

3. Proposed Characters

Annotations in parentheses address special issues for a character, or reference to figures where such special issues are discussed. (These annotations are not intended to be retained in the character list when copied into the standard.)

As the figures are ordered by language, evidence for the characters is found in the figures related to the language, besides any "Yu-x-yy" reference which refers to the specimen numbered "x-yy" in fig. "x".

Languages and other annotations are usually given for the lowercase letters. The alphabets designed for Adyghe and Abkhaz are lowercase-only. Thus, evidence for the capital letters is to be looked at in the figures for other languages.
Proposal to encode Latin letters used in the Former Soviet Union

<table>
<thead>
<tr>
<th>AB3</th>
<th>AB4</th>
<th>AB5</th>
<th>AB6</th>
<th>AB7</th>
<th>AB8</th>
<th>AB9</th>
<th>ABA</th>
<th>ABB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 h</td>
<td>p</td>
<td>x</td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 h</td>
<td>p</td>
<td>x</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 h</td>
<td>p</td>
<td>x</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 h</td>
<td>p</td>
<td>x</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 e</td>
<td>i</td>
<td>s</td>
<td>x</td>
<td>j</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 e</td>
<td>i</td>
<td>s</td>
<td>x</td>
<td>j</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 c</td>
<td>k</td>
<td>s</td>
<td>z</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 c</td>
<td>k</td>
<td>s</td>
<td>z</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 d</td>
<td>l</td>
<td>t</td>
<td>z</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 d</td>
<td>l</td>
<td>t</td>
<td>z</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A t</td>
<td>l</td>
<td>b</td>
<td>z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B g</td>
<td>l</td>
<td>b</td>
<td>z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C g</td>
<td>l</td>
<td>b</td>
<td>z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D g</td>
<td>l</td>
<td>b</td>
<td>z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E q</td>
<td>l</td>
<td>b</td>
<td>z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F q</td>
<td>l</td>
<td>b</td>
<td>z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB0</td>
<td>FB1</td>
<td>FB2</td>
<td>FB3</td>
<td>FB4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>g</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Proposal to encode Latin letters used in the Former Soviet Union**

Date: 19-Oct-2011
**Block: Latin Extended-E**

**Orthographies of the Soviet Union used in the 1920s/1930s**

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c)</td>
<td>LATIN CAPITAL LETTER C WITH MIDDLE TILDE</td>
</tr>
</tbody>
</table>
| \(\epsilon\) | LATIN SMALL LETTER C WITH MIDDLE TILDE  
- Komi |
| \(c\) | LATIN CAPITAL LETTER C WITH DESCENDER  
→ 2185 roman numeral six late form |
| \(\varsigma\) | LATIN SMALL LETTER C WITH DESCENDER  
→ 0481 cyrillic small letter koppa  
→ 04AB cyrillic small letter es with descender  
- Abkhaz, Adyghe, Kabardian, Komi, Udi  
(see fig. Adyghe-1 for glyph with descender at extreme right)  
(see fig. Kabardian-3 for contrastive use to c with cedilla) |

Yu 1-01: already accepted for Unicode 6.1 (U+A792/A793)

Yu 1-02: lowercase is U+0111, uppercase see below as "Alphabetic Presentation Form"

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d)</td>
<td>LATIN CAPITAL LETTER D WITH DESCENDER</td>
</tr>
</tbody>
</table>
| \(\delta\) | LATIN SMALL LETTER D WITH DESCENDER  
- Adyghe, Komi  
(Yu-1-03) |
| \(\iota\) | LATIN SMALL LETTER REVERSED F  
→ A7FB latin epigraphic letter reversed f  
- as capital form, U+A7FB is to be used by tailoring  
- Abaza, Abkhaz, Adyghe  
(Yu-1-04) |

(The case relationship to U+A7FB LATIN EPIGRAPHIC LETTER REVERSED F cannot be reflected in the character properties due to stability policies.)

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
</table>
| \(\gamma\) | LATIN SMALL LETTER SCRIPT G WITH ASCENDER  
- Abkhaz  
(Yu-1-05) |
| \(\Gamma\) | LATIN CAPITAL LETTER SHUGHNI GHE |
| \(\gamma\) | LATIN SMALL LETTER SHUGHNI GHE  
- Shughni |
U+AB6E  LATIN CAPITAL LETTER GHA WITH STROKE THROUGH DESCENDER
U+AB6F  LATIN SMALL LETTER GHA WITH STROKE THROUGH DESCENDER
   → 01A3 latin small letter oi
   •  Tshakur
Yu 1-06: is U+01E5 (Adyghe)
Yu 1-07: lowercase is U+0127, uppercase see below as "Alphabetic Presentation Form"
Yu 1-08: probably is U+2C68 LATIN SMALL LETTER H WITH DESCENDER

U+AB70  LATIN CAPITAL LETTER WIDE H WITH LEFT HOOK
U+AB71  LATIN SMALL LETTER H WITH LEFT HOOK
   •  Abaza, Kabardian
      (Yu-1-09)
Yu 1-10: lowercase is U+0266, uppercase see below as "Alphabetic Presentation Form"
Yu 1-11: lowercase is U+A727, uppercase see below as "Alphabetic Presentation Form"

U+AB72  LATIN CAPITAL LETTER GHE WITH MIDDLE HOOK
U+AB73  LATIN SMALL LETTER GHE WITH MIDDLE HOOK
   → U+0495 cyrillic small letter ghe with middle hook
   •  Kabardian
      (Yu-1-12)

U+AB74  LATIN CAPITAL LETTER I WITH DESCENDER
   → 0131 latin small letter dotless i
U+AB75  LATIN SMALL LETTER DOTLESS I WITH DESCENDER
   •  Azerbaijani, Nogai
      (Yu-2-01)

U+AB76  LATIN CAPITAL LETTER KRA WITH DESCENDER
(see fig. Kabardian-2)
U+AB77  LATIN SMALL LETTER KRA WITH DESCENDER
   → U+0138 latin small letter kra
   •  Abaza, Kabardian
      (see fig. Abaza-4 showing contrastive use to
        U+2C6A LATIN SMALL LETTER K WITH DESCENDER)
L  U+AB78  LATIN CAPITAL LETTER L WITH HOOK
L  U+AB79  LATIN SMALL LETTER WITH HOOK AND TURNED DESCENDER
  • Abaza
    (Yu-2-02)
L  U+AB7A  LATIN CAPITAL LETTER L WITH DESCENDER
L  U+AB7B  LATIN SMALL LETTER L WITH DESCENDER
  • Abaza, Kabardian, Komi
    (Yu-2-03)
L  U+AB7C  LATIN SMALL LETTER L WITH DEFERRED DESCENDER
  • Adyghe, Avar
L  U+AB7D  LATIN SMALL LETTER L WITH TURNED DESCENDER
  • Adyghe
L  U+AB7E  LATIN SMALL LETTER L WITH TURNED DESCENDER AND STROKE
  • Adyghe, Avar
  (the following characters are Yu-2-05, Yu-2-03, Yu-2-04a, reordered to prevent case pairs spreading over columns)

P  U+AB7F  LATIN LETTER SMALL CAPITAL P WITH LOW LEFT HOOK
  → 1D18 latin letter small capital p.
    • Adyghe
      (Yu-2-05)
P  U+AB80  LATIN CAPITAL LETTER OPEN P
(see fig. Kurdish-4)
P  U+AB81  LATIN SMALL LETTER OPEN P
  → A783 latin small letter insular r
    • Abaza, Abkhaz, Tshakur, Udi
      (Yu-2-04)
P  U+AB82  LATIN CAPITAL LETTER P WITH APPENDAGE
P  U+AB83  LATIN SMALL LETTER P WITH APPENDAGE
  → 048F cyrillic small letter er with tick
    • Avar, Kurdish, Lak, Lezgian
    • the appendage may take the shape of a comma or a wavy line
      (Yu-2-04a)
S  U+AB84  LATIN CAPITAL LETTER S WITH SHORT STROKE
S  U+AB85  LATIN SMALL LETTER S WITH SHORT STROKE
  → A7A9 latin small letter s with oblique stroke
  • Komi, Kumyk, Tshakur
  • stroke varies between vertical, diagonal, and horizontal form
    (Yu-2-06)
Yu 2-07: lowercase is U+017F, capital letter see U+ABA9
S  U+AB86  LATIN CAPITAL LETTER S WITH DESCENDER
  → 015E latin capital letter s with cedilla
S  U+AB87  LATIN SMALL LETTER S WITH DESCENDER
  • Abaza, Abkhaz, Avar, Bashkir
T  U+AB88  LATIN CAPITAL LETTER T WITH DESCENDER
t  U+AB89  LATIN SMALL LETTER T WITH DESCENDER
  • Abaza, Abkhaz, Avar, Komi, Kurdish, Tshakur, Udi
    (Yu-2-08)
B  U+AB8A  LATIN CAPITAL LETTER T WITH BOWL
  • glyph variants look like U+0462 cyrillic capital letter yat
b  U+AB8B  LATIN SMALL LETTER T WITH BOWL
  • Adyghe, Bashkir, Shughni
  • glyph variants look like U+0180 latin small letter b with stroke
    (Yu-2-09)
  (the representative glyph for the capital letter is taken from Shugni;
   the one for the small letter is taken from Adyghe)
P  U+AB8C  LATIN CAPITAL LETTER TP
  → 01A4 latin capital letter p with hook
P  U+AB8D  LATIN SMALL LETTER TP
  • Abaza, Abkhaz
    (Yu-2-10)
P  U+AB8E  LATIN CAPITAL LETTER TP WITH DESCENDER
P  U+AB8F  LATIN SMALL LETTER TP WITH DESCENDER
  • Abaza, Abkhaz
    (Yu-2-11)
X U+AB90 LATIN CAPITAL LETTER X WITH DESCENDER
X U+AB91 LATIN SMALL LETTER X WITH DESCENDER
  → 04B3 cyrillic small letter ha with descender
  • Avar, Lak, Tshakur
  (Yu-3-01)
X U+AB92 LATIN CAPITAL LETTER X WITH STROKE
X U+AB93 LATIN SMALL LETTER X WITH STROKE
  → 04FF cyrillic small letter ha with stroke
  • Avar, Lak, Shugni, Tshakur
  (Yu-3-02)
X U+AB94 LATIN CAPITAL LETTER X WITH TWO HOOKS ABOVE
X U+AB95 LATIN SMALL LETTER X WITH TWO HOOKS ABOVE
  • Abaza, Kabardian
  (Yu-3-03)
X U+AB96 LATIN SMALL LETTER SCRIPT X
  • Adyghe
  • typographically a ligature of reversed c and c
Z U+AB97 LATIN CAPITAL LETTER Z WITH STROKE AND DESCENDER
Z U+AB98 LATIN SMALL LETTER Z WITH STROKE AND DESCENDER
  → 01B6 latin small letter z with stroke
  → 26C6 latin small letter z with descender
  • Tshakur
  (Yu-3-04)
3 U+AB99 LATIN CAPITAL LETTER ZE
  → 021C latin capital letter yogh
  → 0417 cyrillic capital letter ze
3 U+AB9A LATIN SMALL LETTER ZE
  → 0292 latin small letter ezh
  • Azerbaijani, Sami, Shugni; 19th century Slovene
  (Yu-3-06; see also fig. 7, 8)
5 U+AB9B LATIN SMALL LETTER LOW TURNED C WITH ASCENDER
  • Adyghe
U+AB9C  LATIN CAPITAL LETTER CAUCASIAN CHE

U+AB9D  LATIN SMALL LETTER CAUCASIAN CHE
→ 04BD cyrillic small letter abkhassian che
• Abaza, Abkhaz, Udi
(Yu-3-07)

U+AB9E  LATIN CAPITAL LETTER CAUCASIAN CHE WITH DESCENDER

U+AB9F  LATIN SMALL LETTER CAUCASIAN CHE WITH DESCENDER
→ 04BF cyrillic small letter abkhassian che with descender
• Abaza, Abkhaz, Udi
(Yu-3-08)

U+ABA0  LATIN LETTER SMALL CAPITAL GHE
→ 0433 cyrillic small letter ghe
→ 1D26 greek letter small capital gamma
• Adyghe
(Yu-3-09)
(a "latin small letter ghe" is presumed to have a typographic descender,
in accordance with Latin che or ze, or the reversed ghe shown below.
As Adyghe makes use of "small capital" letters anyway, the identification
as another "small capital" letter is considered appropriate.)

U+ABA1  LATIN SMALL LETTER REVERSED GHE
→ A780 latin capital letter turned I
• as capital form, U+A780 is to be used by tailoring
• Abaza, Abkhaz, Kabardian, Udi
(Yu-3-10)

U+ABA2  LATIN CAPITAL LETTER REVERSED GHE WITH STROKE
→ A780 latin capital letter turned I

U+ABA3  LATIN SMALL LETTER REVERSED GHE WITH STROKE
• Abaza, Abkhaz, Kabardian, Udi
(Yu-3-11)

U+ABA4  LATIN CAPITAL LETTER REVERSED GHE WITH LOW LEFT HOOK

U+ABA5  LATIN SMALL LETTER REVERSED GHE WITH LOW LEFT HOOK
• Abaza, Abkhaz, Adyghe
(Yu-3-12)
 Proposal to encode Latin letters used in the Former Soviet Union

2011-10-18

Yu 3-14: lowercase is U+0265, uppercase see below as "Alphabetic Presentation Form"

- Adyghe
  (Yu-3-13)

Ü
U+ABA7  LATIN CAPITAL LETTER TURNED WIDE H WITH STROKE THROUGH DESCENDER
→ 0427 cyrillic capital letter che

ü
U+ABA8  LATIN SMALL LETTER TURNED H WITH STROKE THROUGH DESCENDER
→ 0265 latin small letter turned h
  • Abkhaz, Udi
  (Yu-3-15)

Block: Alphabetic Presentation Forms

(The code points are selected to have no overlap with the proposal WG2 N4079, which utilizes the first free codepoints FB07...FB0E following a similar rationale. Also, to have the H-related letters together in a sequence, the gap FB18...FB1C is used, rather than to scatter these over the next free gaps FB10...FB12 and FB18...FB1C.)

Capital letters for orthographies of the Soviet Union used in the 1920s/1930s

The lowercase counterparts of these letters are formally paired with those capital letters to which a compatibility equivalence is declared.

D
U+FB0F  LATIN CAPITAL LETTER D WITH RIGHT STROKE
= latin capital letter Bashkir dhe
→ 0111 latin small letter d with stroke
≈ 0110 latin capital letter d with stroke
  • Bashkir
  (Yu-1-02)

H
U+FB18  LATIN CAPITAL LETTER H WITH LEFT STROKE
→ 0127 latin small letter h with stroke
≈ 0126 latin capital letter h with stroke
  • Avar, Udi
  (Yu-1-07)
Proposal to encode Latin letters used in the Former Soviet Union

Page 16 of 50

2011-10-18

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>U+F819</td>
<td>LATIN CAPITAL LETTER H WITH VERTICAL STROKE</td>
</tr>
<tr>
<td></td>
<td>→ 0127 latin small letter h with stroke</td>
</tr>
<tr>
<td></td>
<td>≈ 0126 latin capital letter h with stroke</td>
</tr>
<tr>
<td></td>
<td>• Kurdish, Judeo-Tat</td>
</tr>
<tr>
<td></td>
<td>(Yu-1-07)</td>
</tr>
<tr>
<td></td>
<td>(for Judeo-Tat, see fig. 8)</td>
</tr>
<tr>
<td>U+F81A</td>
<td>LATIN CAPITAL LETTER WIDE H WITH HOOK</td>
</tr>
<tr>
<td></td>
<td>→ 0266 latin small letter h with hook</td>
</tr>
<tr>
<td></td>
<td>→ 0266 cyrillic capital letter hwe</td>
</tr>
<tr>
<td></td>
<td>≈ A726A latin capital letter h with hook</td>
</tr>
<tr>
<td></td>
<td>• Abaza, Kabardian</td>
</tr>
<tr>
<td></td>
<td>(Yu-1-10)</td>
</tr>
<tr>
<td>U+F81B</td>
<td>LATIN CAPITAL LETTER WIDE HENG</td>
</tr>
<tr>
<td></td>
<td>→ A727 latin small letter heng</td>
</tr>
<tr>
<td></td>
<td>≈ A726 latin capital letter heng</td>
</tr>
<tr>
<td></td>
<td>• Ingush, Nivkh</td>
</tr>
<tr>
<td></td>
<td>(Yu-1-11)</td>
</tr>
<tr>
<td>U+F81C</td>
<td>LATIN CAPITAL LETTER TURNED WIDE H</td>
</tr>
<tr>
<td></td>
<td>≈ latin capital letter che</td>
</tr>
<tr>
<td></td>
<td>→ 0265 latin small letter turned h</td>
</tr>
<tr>
<td></td>
<td>→ 0427 latin capital letter che</td>
</tr>
<tr>
<td></td>
<td>≈ A78D latin capital letter turned h</td>
</tr>
<tr>
<td></td>
<td>• Abaza; 19th century Slovene</td>
</tr>
<tr>
<td></td>
<td>(Yu-3-14; see also fig. 7)</td>
</tr>
</tbody>
</table>

Properties

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB64</td>
<td>LATIN CAPITAL LETTER C WITH MIDDLE TILDE</td>
</tr>
<tr>
<td>AB65</td>
<td>LATIN SMALL LETTER C WITH MIDDLE TILDE</td>
</tr>
<tr>
<td>AB66</td>
<td>LATIN CAPITAL LETTER C WITH DESCENDER</td>
</tr>
<tr>
<td>AB67</td>
<td>LATIN SMALL LETTER C WITH DESCENDER</td>
</tr>
<tr>
<td>AB68</td>
<td>LATIN CAPITAL LETTER D WITH DESCENDER</td>
</tr>
<tr>
<td>AB69</td>
<td>LATIN SMALL LETTER D WITH DESCENDER</td>
</tr>
<tr>
<td>AB70</td>
<td>LATIN CAPITAL LETTER D WITH LEFT HOOK</td>
</tr>
<tr>
<td>AB71</td>
<td>LATIN SMALL LETTER D WITH LEFT HOOK</td>
</tr>
<tr>
<td>AB72</td>
<td>LATIN CAPITAL LETTER H WITH MIDDLE HOOK</td>
</tr>
<tr>
<td>AB73</td>
<td>LATIN SMALL LETTER H WITH MIDDLE HOOK</td>
</tr>
<tr>
<td>AB74</td>
<td>LATIN CAPITAL LETTER I WITH DESCENDER</td>
</tr>
<tr>
<td>AB75</td>
<td>LATIN SMALL LETTER DOTLESS I WITH DESCENDER</td>
</tr>
<tr>
<td>AB76</td>
<td>LATIN CAPITAL LETTER KRA WITH DESCENDER</td>
</tr>
<tr>
<td>AB77</td>
<td>LATIN SMALL LETTER KRA WITH DESCENDER</td>
</tr>
<tr>
<td>AB78</td>
<td>LATIN CAPITAL LETTER L WITH HOOK</td>
</tr>
<tr>
<td>AB79</td>
<td>LATIN SMALL LETTER L WITH HOOK AND TURNED DESCENDER</td>
</tr>
<tr>
<td>AB80</td>
<td>LATIN CAPITAL LETTER L WITH DESCENDER</td>
</tr>
</tbody>
</table>
Proposal to encode Latin letters used in the Former Soviet Union

Page 17 of 50

2011-10-18
4. Acknowledgements

Many thanks to the Russian Wikipedia user "Geoalex" for supplying many valuable information. We respect his wish to not publish his full name here.

Also; special thanks to Juris Cibuls for providing many examples from his collection of historic primers, and for his critical remarks based on them.

5. References

Here, references are listed which are used in the introduction part of this proposal or otherwise generally. References applying to single figures are given in the figure legend.

[1] Юшманов, Николай Владимирович [Yushmanov, Nikolai Vladimirovich]: Определитель Языков [Determinant of languages]. Москва/Ленинград [Moscov/Leningrad (now Saint Petersburg); Russia] 1941

[2] Марр, Николай Яковлевич [Marr, Nikolai Yakovlevich]: Абхазский аналитический алфавит [Abkhaz analytical alphabet]. Ленинград [Leningrad, now Saint Petersburg; Russia] 1926


[7] Russian Wikipedia user "Geoalex", http://ru.wikipedia.org/wiki/%D0%A3%D1%87%D0%B0%D1%81%D1%82%D0%BD%D0%B8%D0%BA:Geoalex (retrieved 2011-10-02). See also section 4. "Acknowledgements".

6. Examples and Figures

Figures which apply to several languages, or which provide additional evidence outside the scope of this proposal for characters proposed here, are given decimal numbers and are listed first.

They are followed by figures which apply to a specific language. These are identified by the English language name + "," + a decimal number. They are listed alphabetically by language name, and within a language section by the number part.

If a specimen showing the complete alphabet of a language was found and considered relevant, usually it is given as the first figure filed for this language (thus numbered "Language"-1).

Characters proposed in this document are referenced here by their name, without the proposed code point. Already encoded letters are referenced by their code point (U+xxxx) and name.

Disclaimer: The political, ideological, philosophical, or religious views expressed in some of the text specimens presented here are not necessarily the ones of the authors of this proposal.
Fig. 1: From [1], p. 28/29. 
Showing the first column, with the references added which are used in this proposal.

ä вепс., икор., ингуш., карел.  
Ʌ адыгейский  
æ ингуш., кетский, осет., селькупский, удэйский  
Ç ингуш., осет., чеченский  
1-01 ç курдский, саамский  
Ç [большинство¹]  
1-02 á абхаз., адых., башк., шугн.  
1-03 å адых., белуджский  
æ [большинство²]  
ê курдский  
1-04 í абаз., абхаз., адых., кабард.  
1-05 ô абазинский, абхазский  
1-06 ñ адыхейский  
ouns [большинство³]  
1-07 ö абаз., адых., дагест., курд., татск.  
1-08 ô дагест., татск., шугнанский  
1-09 ô абаз., кабардинский  
1-10 ô абаз., адых., кабардинский  
1-11 ô ингуш., кабардинский, кетский, маньс., ненецкий, нивх., осет.  
1-12 ō адыхейский  
ı абазинский, абхазский  
î таджикский
Fig. 2: From [1], p. 28/29.
Showing the second column, with the references added which are used in this proposal.

<table>
<thead>
<tr>
<th>Proposal</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-01</td>
<td>хакасский</td>
</tr>
<tr>
<td></td>
<td>тувинский</td>
</tr>
<tr>
<td></td>
<td>кавказские, курдский</td>
</tr>
<tr>
<td></td>
<td>к адыгейский</td>
</tr>
<tr>
<td>2-02</td>
<td>адыг., кабардинский</td>
</tr>
<tr>
<td></td>
<td>авар., ительм., хант., эским.</td>
</tr>
<tr>
<td>2-03</td>
<td>абаз., авар., адыг., кабард.</td>
</tr>
<tr>
<td></td>
<td>хантыйский</td>
</tr>
<tr>
<td></td>
<td>ингуш., тюркские, чеченский</td>
</tr>
<tr>
<td></td>
<td>языки народов Севера</td>
</tr>
<tr>
<td></td>
<td>вепс., ижор., карел., чеченский</td>
</tr>
<tr>
<td></td>
<td>[большинство]</td>
</tr>
<tr>
<td></td>
<td>адыгейский</td>
</tr>
<tr>
<td>2-04</td>
<td>абаз., абхаз., кабард.</td>
</tr>
<tr>
<td>2-04a</td>
<td>дагест. (кроме авар.), курд.</td>
</tr>
<tr>
<td>2-05</td>
<td>адыгейский</td>
</tr>
<tr>
<td></td>
<td>абаз., абхаз., дагест., кабард.</td>
</tr>
<tr>
<td></td>
<td>нивхский</td>
</tr>
<tr>
<td></td>
<td>адыг., ингуш., чеченский</td>
</tr>
<tr>
<td></td>
<td>[большинство]</td>
</tr>
<tr>
<td>2-06</td>
<td>дагест., маньс., саам., хант.</td>
</tr>
<tr>
<td>2-07</td>
<td>абазинский, абхазский</td>
</tr>
<tr>
<td>2-08</td>
<td>белудж., кавк., курд., сир.</td>
</tr>
<tr>
<td>2-09</td>
<td>адыг., башк., шугнанский</td>
</tr>
<tr>
<td>2-10</td>
<td>абхазский</td>
</tr>
<tr>
<td></td>
<td>2-11</td>
</tr>
</tbody>
</table>
Fig. 3: From [1], p. 28/29.  
Showing the third column, with the references added which are used in this proposal.

ё чеченский
ё см. ё,  ý
ё казах., каракалпакский
ё таджикский
х ингуш., чеченский

3-01 х дагестанские
3-02 х дагест., шугнанский
3-03 х абаз., адиг., кабард.
3-04 х ингуш., осет., чеченский
3-05 х кавказские
3-06 х дунганский
3-07 х [большинство 7]
3-08 х даргинский, табасаранский
3-09 х кавказские, нанайский, саам., уде́йск., эвенк., эвенск.
3-10 х коми
3-11 х саамский, шугнанский
3-12 х абазинский, абхазский
3-13 х [большинство 8]
3-14 х адигейский
3-15 х абаз., абхаз., кабард.

в цел., таджикский, тюркские

(реже другие).
Proposal to encode Latin letters used in the Former Soviet Union

2011-10-18
Fig. 5: [2], vol. 3, p. 134.

Showing the Latin and Arabic alphabets for Avar, Dargwa, Lak, and Lezgian.

<table>
<thead>
<tr>
<th>Avarian</th>
<th>Darginian</th>
<th>Lakian</th>
<th>Lezgian</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
<td>e</td>
<td>e</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>f</td>
<td>f</td>
</tr>
<tr>
<td>g</td>
<td>g</td>
<td>g</td>
<td>g</td>
</tr>
<tr>
<td>h</td>
<td>h</td>
<td>h</td>
<td>h</td>
</tr>
<tr>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
</tr>
<tr>
<td>k</td>
<td>k</td>
<td>k</td>
<td>k</td>
</tr>
<tr>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
</tr>
<tr>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
</tr>
<tr>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>q</td>
<td>q</td>
<td>q</td>
<td>q</td>
</tr>
<tr>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
</tr>
<tr>
<td>s</td>
<td>s</td>
<td>s</td>
<td>s</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>u</td>
<td>u</td>
<td>u</td>
<td>u</td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>z</td>
<td>z</td>
<td>z</td>
<td>z</td>
</tr>
</tbody>
</table>
Proposal to encode Latin letters used in the Former Soviet Union

Showing additional evidence for LATIN LETTER H WITH VERTICAL STROKE for an application related to, but not exactly within the scope of this proposal.

This is Fig. 1 from L2/08-034R (submitted 2008-04-18 by Lorna Priest).
Showing the Metelko alphabet (which was used for the Slovene language in the 1820s) in relation to contemporary Latin alphabets (referred as "Neu-Krain.", i.e. "new Slovene"). While this alphabet itself is not subject of this proposal, it contains letters which are used in the Latin alphabets of the Soviet Union also about 100 years later, and therefore are proposed here. A feature of this alphabet is that letters which are modeled after Cyrillic letters are adapted to have a real Latin look. Regarding the Latin small ZE and small CHE, this is done by expanding the letter downwards, giving them a true x-height body and a full typographic descender. In fact, the small CHE can be unified with U+0265 LATIN SMALL LETTER TURNED H.

Fig. 7: [6], p. 1:
Text sample of the Metelko alphabet used in Slovenia in the 1820s, showing LATIN CAPITAL/SMALL LETTER ZE, the latter as a true Latin form with x-height body and typographic descender.

Fig. 8: Excerpt from [6], p. 281: Text sample of the Metelko alphabet used in Slovenia in the 1820s, showing LATIN CAPITAL/SMALL LETTER ZE, the latter as a true Latin form with x-height body and typographic descender.
| Aa | Ab | Ñë | ÑÔ | Tž | Tť | Žàž | Ƣğ | Ƣụ | Ƣї | Ƣї | Kļ | Kļ |
| Bb | Bb | Íí | Íí | Tř | Tř | Žř | Ƣř | Ƣř | Ƣř | Ƣř | Fī | Fī |
| Cc | Çc | Ňč | Ňč | Óo | Óo | Ùu | Ùu | Ùu | Ùu | Fb | Fb |
| Dd | Ðd | Êë | Êë | Íñ | Íñ | Úx | Úx | Ùx | Ùx | Hê | Hê |
| Èe | Èc | Êê | Êê | Òo | Òo | Ñx | Ñx | Ñx | Ñx | Uê | Uê |
| FF | FF | Čj | Čj | Žj | Žj | Íł | Íł | Íł | Íł | Kî | Kî |
| Ңң | Ңң | Kk | Kk | Rr | Rr | Pp | Uu | Þţ | Þţ | Гг | Гг |
| Qq | Qq | Ёж | Ёж | Сс | Сс | Zz | Zz | Zz | Zz | Кк | Кк |
| Ss | Ss | Žš | Žš | Ñž | Ñž | Kk | Kk | Kk | Kk | Ёж | Ёж |
| g̣ | g̣ | Ѧђ | Ѧђ | ѧѣ | ѧѣ | Ѩѣ | Ѩѣ | ѩѣ | ѩѣ | Фф | Гг |

**Fig. Abaza-1:** Correspondence between the historic Latin and the current Cyrillic orthography for Abaza, compiled by "Geoalex" [7].

Retrieved 2010-05-12 from: http://ru.wikipedia.org/wiki/%D0%A4%D0%B0%D0%B9%D0%BB:Abaza_latin_%26_cyrillic_alphabets.jpg

**Fig. Abaza-2:** Title page and sample page from a book published 1933.
Fig. Abaza-4: Excerpt of the sample page shown in fig. Abaza-2.
Marked red: LATIN CAPITAL LETTER X WITH TWO HOOKS
Marked green: LATIN CAPITAL LETTER TURNED WIDE H ("Latin che")
Marked blue: LATIN CAPITAL LETTER S WITH DESCENDER

Fig. Abaza-4: ibid., p. 3, excerpt of the sample page shown in fig. Abaza-2.
Marked red: LATIN CAPITAL LETTER WIDE H WITH HOOK as uppercase counterpart to U+0266 LATIN SMALL LETTER H WITH HOOK.
Marked green: LATIN SMALL LETTER KRA WITH DESCENDER (last line) in contrast to U+2C6A LATIN SMALL LETTER K WITH DESCENDER (second line).
Marked blue: LATIN SMALL LETTER L WITH HOOK AND TURNED DESCENDER. The existence of the uppercase counterpart is deduced, as Abaza uses capital letters.
Marked purple: LATIN SMALL LETTER H WITH LEFT HOOK.
Marked orange: LATIN SMALL LETTER CAUCASIAN CHE.
Marked grey: LATIN SMALL LETTER S WITH DESCENDER.

Fig. Abaza-5: ibid., p. 5.
Marked red: LATIN SMALL LETTER REVERSED GHE WITH STROKE (here with a glyph variant without a typographical descender)
Marked green: LATIN SMALL LETTER X WITH TWO HOOKS.

Fig. Abaza-6: ibid., p. 60, showing LATIN CAPITAL LETTER WIDE H WITH LEFT HOOK, LATIN SMALL LETTER H WITH LEFT HOOK.
Marked red: x-height version of U+01A3 LATIN SMALL LETTER OI (a.k.a. GHA) considered as glyph variation.  
Marked green: LATIN CAPITAL LETTER TP WITH DESCENDER  
Marked blue: LATIN SMALL LETTER C/S WITH DESCENDER, showing examples where the descender is applied right from the center  
Marked grey: U+0265 LATIN SMALL LETTER TURNED H  
Marked purple: LATIN CAPITAL/SMALL LETTER OPEN P  
Marked brown: U+A7FB LATIN EPIGRAPHIC LETTER REVERSED F and LATIN SMALL LETTER REVERSED F as case pair  
Marked light blue: LATIN CAPITAL LETTER REVERSED GHE WITH LEFT HOOK  
Marked orange: LATIN CAPITAL LETTER TURNED WIDE H WITH STROKE THROUGH DESCENDER; LATIN SMALL LETTER TURNED H WITH STROKE THROUGH DESCENDER

---

Fig. Abaza-8: ibid., p.15 – Marked red:  
LATIN CAPITAL LETTER REVERSED GHE WITH STROKE.
Fig. Abkhaz-1: Title page and sample page of a book published 1935. Also, an enlarged excerpt of the sample page, showing the complete (lowercase-only) Abkhaz alphabet.

Fig. Abkhaz-2: Correspondence between the historic Latin and the current Cyrillic orthography for Adyghe, based on a table uploaded by Yuri Koryakov to the Russian Wikipedia, retrieved 2010-11-20 from: http://ru.wikipedia.org/wiki/%D0%A4%D0%B0%D0%BD%D0%BC:Abxaz_lat%2BIPA.png
Fig. Abkhaz-3: Title page and sample page from two books published 1934 and 1937. The first one is a collection of poems; the second one is a translation of a work from the Spanish author Lope de Vega (1562-1635).
Fig. Adyghe-1: From a book published in 1927 introducing the Adyghe alphabet: Title, sample page, and the two pages showing the complete (lowercase only) alphabet. Note especially on the alphabet pages: C WITH DESCENDER (left page, first row, 4th char) having the descender rightmost. KRA WITH DESCENDER having a right top different from K (left page, 4th and 5th row).
Fig. Adyghe-2: Correspondence between the historic Latin and the current Cyrillic orthography for Adyghe, compiled by "Geoalex" [7].
Retrieved 2010-07-23 from:
http://ru.wikipedia.org/wiki/%D0%A4%D0%B0%D0%BF%D0%B9%D0%BB:Adyghe_alphabet_cyrillic-latin.JPG

Fig. Adyghe-3: Title page and sample page from a book published 1931.
Fig. Adyghe-4: Showing the cover of a book published 1932 which shows the slogan “Workers of the world, unite!” in several languages. The first line is Adyghe, showing its lowercase-only orthography. Alto, the inner title is shown (which contains the reference in Russian and Adyghe).
Avar: see also fig. 5

**Fig. Avar-1:** Title and sample page of a book published in 1935, and an enlarged excerpt of p.13. Marked red: LATIN CAPITAL LETTER L WITH DESCENDER, LATIN CAPITAL/SMALL LETTER T WITH DESCENDER, LATIN CAPITAL/SMALL LETTER X WITH DESCENDER
Marked green: LATIN SMALL LETTER L WITH DEFERRED DESCENDER
Marked orange: LATIN SMALL LETTER L WITH DEFERRED DESCENDER AND STROKE
Marked blue: LATIN CAPITAL/SMALL LETTER S WITH SHORT STROKE
(showing a glyph with diagonal stroke for the capital letter, but with horizontal stroke for the small letter)
Marked purple: LATIN CAPITAL LETTER H WITH LEFT STROKE
Marked brown: LATIN CAPITAL/SMALL LETTER X WITH STROKE
**Fig. Azerbaijani-1:**

Shown as fig. 43 on p.367 in [5].
*Showing LATIN CAPITAL LETTER I WITH DESCENDER and LATIN SMALL LETTER DOTLESS I WITH DESCENDER (red); showing LATIN CAPITAL LETTER ZE (green).*

<table>
<thead>
<tr>
<th>A a</th>
<th>B b</th>
<th>C c</th>
<th>D d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>E e</td>
<td>F f</td>
<td>G g</td>
<td>H h</td>
</tr>
<tr>
<td>i</td>
<td>j</td>
<td>k</td>
<td>l</td>
</tr>
<tr>
<td>m</td>
<td>n</td>
<td>o</td>
<td>p</td>
</tr>
<tr>
<td>q</td>
<td>r</td>
<td>s</td>
<td>t</td>
</tr>
<tr>
<td>u</td>
<td>v</td>
<td>x</td>
<td>z</td>
</tr>
</tbody>
</table>

Fig. Azerbaijani-2: ibid., p. 35. Shown as fig. 45 on p.376 in [5].
*Showing the letters marked in fig. Azerbaijani-1 by the same colors.*

Ferc: Cefer: bir de Befike dəftərlerində seçilən cəqirdilər. Ferc dedi: „Men dəf seçilən cəqirmə; Cefer, sen fəs seçilən cəq! Befike de qəsəq qəsən, qəreq, qim jaxə, qəreq Əçədilər, kyrtdilər. Anəları, Fatma seçilləre baxib xəsələb.“
Fig. Bashkir-1: Two pages of a book printed in Bashkir Latin orthography, showing the complete alphabet containing LATIN CAPITAL LETTER D WITH RIGHT STROKE and LATIN CAPITAL/SMALL LETTER T WITH BOWL.

Fig. Bashkir-2: Specimen of text printed in Bashkir Latin orthography, retrieved 2008-10-24 from the Russian Wikipedia at http://ru.wikipedia.org/wiki/%D0%A4%D0%B0%D0%BB:Bashqortalifba.jpg
Marked red: LATIN SMALL LETTER T WITH BOWL.
Fig. Ingush-1: Excerpt from p.4 of: Serdalo (an Ingush magazine), № 2–(10), Vladikavkaz 1924-03-08, explaining the Ingush alphabet.
Fig. Ingush-2: ibid., excerpts from p.2 and p.4, showing LATIN CAPITAL LETTER WIDE HENG and U+A727 LATIN SMALL LETTER HENG in different fonts.

Fig. Ingush-3: Header of a later issue of the same magazine (1926-01-20, then renamed to “Serdælo”). also showing LATIN CAPITAL LETTER WIDE HENG.

Fig. Kabardian-1: Title page and sample page of a book printed 1926. Marked red: LATIN CAPITAL/SMALL LETTER GHE WITH MIDDLE HOOK Marked green: several letters with descender, showing all the same shape. Note that the descenders on the “ı” and the “ż” are centered horizontally, like for the “ç” and the “ś”.

Proposal to encode Latin letters used in the Former Soviet Union 2011-10-18 Page 38 of 50
Fig. Kabardian-2: ibid., p.6:
Marked red: showing LATIN CAPITAL LETTER KRA WITH DESCENDER (second line), also proving that the form of its right top is in fact different from U+004B LATIN CAPITAL LETTER K (second last line).
Marked blue: This small form is considered a glyph variant of U+0266 LATIN SMALL LETTER H WITH HOOK.

Fig. Kabardian-3: ibid., p.51:
Showing SMALL C WITH DESCENDER (green) in contrast to C WITH CEDILLA (red). The use of the C WITH CEDILLA obviously is a typo. However, the existence of different glyphs for these two letters in an original metal type font proves that the letters in fact were regarded as different by the original users of the Latin orthography of Kabardian.

Fig. Kazhak-1: Title page from a Kazhak newspaper from about 1937, showing all proposed letters. Retrieved 2008-10-25 from http://en.wikipedia.org/wiki/Image:Sotsijaldy_qazaqstan.jpg
Shows examples of the Latin Ь/ь which is not proposed here but discussed in the introduction.

Фиг. Кomi-1: Таблица алфавита коми языка 1930-х годов,

Fig. Komi-2: Title page and excerpt from p.17 of a Komi grammar printed in 1933. The excerpt shows LATIN CAPITAL/SMALL LETTER C WITH MIDDLE TILDE.

Фиг. Кomi-2: Титульный лист и отрывок с стр. 17 кн. грамматики на комском языке, вышедшей в 1933 году. В отрывке показаны латинские буквы C и Ŋ.
Fig. Komi-3: Showing the outer and inner title page of a book printed 1933. While both pages share the same content, different fonts are used. This proves the existence of considerable glyph variation regarding the positioning and shape of the descenders and the slanting angle of the stroke of LATIN CAPITAL LETTER WITH SHORT STROKE.

On the outer title page shown left, U+0044 LATIN CAPITAL LETTER D shows a mirrored glyph variant, similar to that shown in fig. Udi-1.

Fig. Komi-4: Ibid., two excerpts from p.51 showing LATIN CAPITAL LETTER S WITH SHORT STROKE, having a vertical stroke in a bold font and a slanted stroke in a common font.
Fig. Komi-5: Title page and sample page (p.85) of a Komi-Russian dictionary printed 1934. Note the glyph variation for LATIN CAPITAL LETTER L WITH DESCENDER on the sample page. Also, "Medeсаь" contains a LATIN SMALL LETTER C WITH MIDDLE TILDE. On the title page, the subtitle starts with LATIN CAPITAL LETTER D WITH DESCENDER.

Fig. Komi-6: showing a song text by the Komi poet Victor Savin (Нёбдінса Виттор; Виктор Алексеевич Савин; 1888–1943). Note that all descenders are placed rightmost here. Retrieved 2011-09-28 from http://foto11.com/komi/art/singing/songsavin.shtml#more
Note the LATIN SMALL LETTER S WITH SHORT STROKE having different orientations of the stroke within the same book. Comparing with fig. Komi-4, it seems that oblique strokes in bold fonts are avoided as they yield in a too "closed" or "8-like" look of the "S".

Fig. Kurdish-1: Title page and sample page of a book printed 1931 in Jerevan (now Armenia), showing the Kurdish alphabet on the sample page.
Marked red: LATIN CAPITAL/SMALL LETTER P WITH APPENDAGE.
Marked green: LATIN CAPITAL LETTER H WITH VERTICAL STROKE.

Fig. Kurdish-3: ibid., enlarged excerpt from p.36.
Marked green: LATIN CAPITAL LETTER H WITH VERTICAL STROKE occurring as uppercase counterpart of U+0127 LATIN SMALL LETTER H WITH STROKE.

Fig. Kurdish-4: ibid., enlarged excerpt from p.39.
Marked red: LATIN CAPITAL/SMALL LETTER P WITH APPENDAGE.

Lak: see also fig. 5

Fig. Lak-1: Excerpt from a report on the Soviet constitution, printed 1937.
Marked red: LATIN SMALL LETTER P WITH APPENDAGE.
Fig. Nivkh-1: showing the outer and inner title of a primer published in 1936. The capitalized title on the inner page shows a glyph variant of LATIN CAPITAL LETTER WIDE HENG.

Fig. Shughni-1: Title page, p.83, and p.105 from a book printed 1931 in Tashkent, containing:
- LATIN CAPITAL/SMALL LETTER SHUGHNI GHE (marked red, common "g" marked blue)
- LATIN CAPITAL/SMALL LETTER T WITH BOWL (marked green)
The alphabet on the bottom of p.16 shows e.g. LATIN CAPITAL/SMALL LETTER GHA WITH STROKE THROUGH DESCENDER, LATIN CAPITAL/SMALL LETTER Z WITH STROKE AND DESCENDER.

(Regarding the CYRILLIC-SHA-like letter, see fig. Tshakur-2 below.)

Fig. Tshakur-1: p. 12 and p. 16 of: Genqo A.: Saxəlfəsəs və jazə qəjdaları. Bakı (Baku), 1934.

The examples given for the CYRILLIC-SHA-like letter in the last excerpt show that it seems to be a vowel. Therefore, we consider it as a glyph variant of U+019C/U+026F LATIN CAPITAL/SMALL LETTER TURNED M, rather than proposing it as a new LATIN CAPITAL/SMALL LETTER SHA. The italic form is compatible with this assumption.
Fig. Udi-1: Title and sample pages (59, 48/49) from a book printed 1934.
(The title page shows a mirrored glyph variant of U+0044 LATIN CAPITAL LETTER D.)
**Fig. Udi-2:** ibid., p.36 and p.43.

Top of p.36, introducing the letter LATIN CAPITAL/SMALL C WITH DESCENDER, using glyphs which show the descender clearly right from the center line.

Bottom of p.36, introducing LATIN CAPITAL LETTER LONG S together with U+017F LATINSMALL LETTER LONG S.

Also, showing LATIN CAPITAL LETTER H WITH VERTICAL STROKE.

p.43, introducing LATIN CAPITAL/SMALL LETTER REVERSED GHE WITH STROKE.
Proposal to encode Latin letters used in the Former Soviet Union

**A. Administrative**

1. **Title:** Proposal to encode Latin letters used in the Former Soviet Union
2. Requester's name: Nurlan Joomageldinov, Karl Pentzlin, Ilya Yevlampiev
3. Requester type (Member body/Liaison/Individual contribution): Expert Contribution
4. Submission date: 2011-10-18
5. Requester's reference (if applicable): 
6. Choose one of the following:
   - This is a complete proposal: Yes
   - (or) More information will be provided later: 

**B. Technical – General**

1. Choose one of the following:
   - a. This proposal is for a new script (set of characters): No
   - b. The proposal is for addition of character(s) to an existing block: Yes
     - Name of the existing block: Latin Extended-E; Alphabetic Presentation Forms
2. Number of characters in proposal: 76
3. Proposed category (select one from below - see section 2.2 of P&P document):
   - A-Contemporary
   - B.1-Specialized (small collection)
   - B.2-Specialized (large collection)
   - C-Major extinct
   - D-Attested extinct
   - E-Minor extinct
   - F-Archaic Hieroglyphic or Ideographic
   - G-Obscure or questionable usage symbols
4. Is a repertoire including character names provided? Yes
   - a. If YES, are the names in accordance with the “character naming guidelines” in Annex L of P&P document? Yes
   - b. Are the character shapes attached in a legible form suitable for review? Yes
5. Fonts related:
   - a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard? The co-author Karl Pentzlin (on request)
   - b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.): The co-author Karl Pentzlin
6. References:
   - a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? Yes
   - b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? Yes
7. Special encoding issues:
   - Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? Yes
   - Special casing issues (see text)
8. Additional Information:
   - Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

---

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? No
   If YES explain

2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? Yes
   If YES, with whom? Linguists, Librarians
   If YES, available relevant documents: see text

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? Yes
   Reference: once used by several millions of people (see text)

4. The context of use for the proposed characters (type of use; common or rare) Historically common
   Reference: see text

5. Are the proposed characters in current use by the user community? Yes
   If YES, where? Reference: see text (e.g. historians, librarians, linguists)
   After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? Yes
   If YES, is a rationale provided? Yes
   If YES, reference: To keep them in line with already encoded characters

6. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? Yes

7. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? No
   If YES, is a rationale for its inclusion provided? Yes
   If YES, reference: (the block “Alphabetic Presentation Forms” is used for other reasons, see text)

8. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? No
   If YES, is a rationale for its inclusion provided? Yes
   If YES, reference: (besides the long-standing fact that there are Cyrillic letters similar to Latin ones)

9. Can any of the proposed characters be considered to be similar (in appearance or function) to an existing character? No
   If YES, is a rationale for its inclusion provided? Yes
   If YES, reference: (besides the long-standing fact that there are Cyrillic letters similar to Latin ones)

10. Does the proposal include use of combining characters and/or use of composite sequences? No
    If YES, is a rationale for such use provided? Yes
    If YES, reference:

11. Does the proposal contain characters with any special properties such as control function or similar semantics? No
    If YES, describe in detail (include attachment if necessary)

12. Does the proposal contain any ideographic compatibility characters? No
    If YES, are the equivalent corresponding unified ideographic characters identified? Yes
    If YES, reference: