Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международнаяорганизацияпостандартизации

Doc Type: Working Group Document
Title: Revised Proposal to encode Arabic characters used for Bashkir, Belarusian, Crimean Tatar, and Tatar languages
Source: See References
Authors: Ilya Yevlampiev, Karl Pentzlin, Nurlan Joomagueldinov
Status: Expert Contribution
Action: For consideration by JTC1/SC2/WG2 and UTC
Date: 20 May 2011
Supersedes: L2/11-138 "Proposal to encode Arabic characters used for Bashkir, Belarusian, Crimean Tatar, Karachay, Karakalpak, and Tatar languages"

This revision incorporates the decisions of UTC #127, as documented in WG2 N4065 = L2/11-201 »Proposed codepoints and properties for characters in"Proposal to encode Arabic characters used for Bashkir, Belarusian, Crimean Tatar, Karachay, Karakalpak, and Tatar languages"«.
The three characters proposed in L2/11-138 which were not accepted by UTC #127 are no longer subject of this proposal. Therefore, the terms "Karachay" and "Karalkapak" were dropped from the title.

1. Introduction

The letters of Bashkir, Belarusian, Crimean Tatar, and Tatar Arabic orthographies that are not encoded in Unicode standard are presented here.

2. Proposed Characters

Annotations below 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.)

Proposed codepoint	Glyph	F M I	Proposed name	Linguistic comment
08AD	١	٠٠١	ARABIC LETTER LOW ALEF • Bashkir, Tatar	*denotes the usage of back vowels in the word with dubious spelling in Yaña imlâ orthography*
08AE	ڈ	ڈ٠٠	ARABIC LETTER DAL WITH THREE DOTS BELOW • Belarusian	*denotes the sound [dz]*
08AF	ڜ	ڜڜڜ	ARABIC LETTER SAD WITH THREE DOTS BELOW • Belarusian	*denotes the sound [ts]*
08B0	ک	ککک	ARABIC LETTER GAF WITH INVERTED STROKE • Crimean Tatar, Chechen, Lak	*denotes the sound [j] in Crimean Tatar* *denotes the sound [kʼ] in Chechen and Lak*
08B1	۶	۶٠٠	ARABIC LETTER STRAIGHT WAW • Tatar	*denotes the sound [ɯ] in Urta imlâ Alparov's orthograpthy*

3. Encoding Considerations

Joining type and group for ArabicShaping.txt:

```
08AD; LOW ALEF; U; No_Joining_Group
08AE; DAL WITH 3 DOTS BELOW; R; DAL
08AF; SAD WITH 3 DOTS BELOW; D; SAD
08B0; KEHEH WITH STROKE BELOW; D; GAF
08B1; STRAIGHT WAW; R; STRAIGHT WAW
```

Unicode character Properties:

```
08AD;ARABIC LETTER LOW ALEF;Lo;0;AL;;;;;N;;;;;
08AE;ARABIC LETTER DAL WITH THREE DOTS BELOW;Lo;0;AL;;;;;N;;;;;
08AF;ARABIC LETTER SAD WITH THREE DOTS BELOW;Lo;0;AL;;;;;N;;;;;
08B0;ARABIC LETTER GAF WITH INVERTED STROKE;Lo;0;AL;;;;;N;;;;;
08B1;ARABIC LETTER STRAIGHT WAW;Lo;0;AL;;;;;N;;;;;
```

Collation:

Characters may be sorted after similar characters.

- U+08AD ARABIC LETTER LOW ALEF should have secondary or tertiary sorting after (its appearance before a word should not affect sorting, it depends on the letters after (U+08AD isn't a letter, but a sign before them). Only in the case when there are two words that totally coincide with each other with all letters and differs only where one has a U+08AD in the beginning and the second does not should the first word be the one with U+08AD (low alef) before it and the word without it be sorted afterward the second.
- U+08AE ARABIC LETTER DAL WITH THREE DOTS BELOW should sort after U+068D ARABIC LETTER DDAHAL
- U+08AF ARABIC LETTER SAD WITH THREE DOTS BELOW should sort afterU+069D ARABIC LETTER SAD WITH TWO DOTS BELOW
- U+08B0 ARABIC LETTER GAF WITH INVERTED STROKE should sort afterU+06AF ARABIC LETTER GAF
- U+08B1 ARABIC LETTER STRAIGHT WAW should sort after U+06CB ARABIC LETTER VE

Confusability

Confusability issues would only arise if any of these characters were decomposed.

4. Usage and Encoding Rationale

Additions for Bashkir, Tatar

The proposed letters were used in 1910s-1920s – in1920s in Yaña imlâ (New Orthography) orthography and several experimental transition orthographies of 1910s colloquially named *Urta imlâ* (Middle Orthography) [1][2][4][5].

For the Tatar the majority of examples are stored here:

http://fotki.yandex.ru/users/ievlampiev/album/64267

http://fotki.yandex.ru/users/ievlampiev/album/67014/ (also Bashkir)

http://fotki.yandex.ru/users/ievlampiev/album/116321/

See fig. 1 for Tatar alphabet.

Explanation of the Bashkir-Tatar Low Alef (based on ref [1][2]): Low Alef (ٳ; Tatar: qalınlıq bilgäse,калынлыкбилгәсе, i.e. the hard sign) is a special character marking, that a word (or at least its first part) have back vowels. It was placed in the beginning of the words only.

Low alef isn't connected to other letters and occurs in initial and standalone positions only.

It works this way: the Tatar language has 5 pairs of vowels, one of the pair mates is a diphthong: ı - e [ы- э], i - ıy (í) [и - ый], o-ö [o - ө], u-ü [y - ү], a-ä [a - ə].In the Yaña imlâ alphabet, all of them used one vowel sign for each pair, except the last,i.e. Farsi yeh-dammaFarsi yeh U+06CC (ۍ), waw-damma U+06C7 (ۇ), waw U+0648 (و), and paired alef U+0627 (ا) and Arabic ae U+06D5 (ە). Possibly, the last pair was signed by different characters due to many Arabic words that don't obey the vowel harmony law.Low alef was used in places where dubious reading might occur, for example, tor (stay) and tör (variety) without it. It was placed before the word that consisted of back vowels, i.e. in "tor" (تورٳ) but not for "tör" (تور). See fig 3.

In some cases it was omitted: for words containing "q" (ق) or "ğ" (ع) - they usually haveback vowels, or words, containing "a"(ا), as their other vowels are back due to vowel harmony law.

Rationale for encoding low alef as a separate character can be derived from 1) its form; 2) its usage.

1) Unlike the subscript alef (U+0656) low alef isn't placed below base characters, but is a standalone alef placed as subscript index. Moreover, subscript alef and low alef coexisted in the Tatar writing, but in the different orthographies (see fig. 4 &5). Low alef appears in the last Arabic orthography Yaña imlâ, and the subscript alef appears in the transition Urta imlâ experimental orthographies, developed and used by many Tatar scholars until the Yaña imlâ was adopted. Subsequently they both should occur in the same font where they will have different glyph view. Also, subscript alef couldn't be used instead of low alef, as in any ordinary font it will be shown as haraka – for historical usage it is better for reader to see the "box" than incorrect view.
2) Low alef has the similar behavior to high hamza (U+0674) used for Arabic in Central Asian languages. It occurs only in beginning of the word and doesn't interacts with other letters of the word. Low alef is not haraka, and not a letter, it has another sense, similar to Central Asian high hamza. See Fig. 3, 9.

Additions for Crimean Tatar, Chechen and Lak

The proposed letter was used in Crimean Tatar to designate [j] (fig. 14) in some orthoraphies and for Chechen and Lak to designate [k'] (fig 15, 16) in 1920s.

For Crimean Tatar usage examples see http://fotki.yandex.ru/users/ievlampiev/album/67179/ .

Additions for Belarusian

The proposed letters were used in handwriting in the 16th-19th centuries for writing Belarusian and Polish among the Belarusian Tatars. In the 20th century it was used in printing among scholars studying this alphabet [3]. See fig 7, 10.

Examples of texts are found at:

http://fotki.yandex.ru/users/ievlampiev/album/64978

5. Acknowledgements

Many thanks to Tatarstan National Library workers for access for copying some texts.
Special thanks to Lorna Priest for valuable help in composing this proposal.

6. References

[1] *See http://fotki.yandex.ru/users/ievlampiev/album/116321/ (2nd and 3rd lines form title an http://fotki.yandex.ru/users/ievlampiev/view/403389?page=0Русьмактәбләрәүчүнтатартелэдәрслеги, Казань, 1925)*

[2]Bashkir grammar (p .III; p. 9) found on: http://fotki.yandex.ru/users/ievlampiev/album/67014/

[3] Антонович А.К. Белорусские тексты, писанныеарабскимписьмом, и их графико-орфографическая система. Вильнюс, 1968. http://fotki.yandex.ru/users/ievlampiev/album/64978/

[4] Курбатов, Хәлиф Рәхим улы. Татар теленең алфавит һәм орфография тарихы / Х.Р. Курбатов.—Казан: Татарстан китап нәшр., 1960.—131 б.; 23 .— <URL:http://z3950.ksu.ru/knigi/X_Qurbatov_Tatar_teleneng_alfavit_ham_orfografia_tarixi.djvu>.

[5] Әхмәров К. З. Башҡорт яҙыуы тарихенән = Из истории башкирской письменности: (башҡорт әҙәби теленең алфавите һәм орфографияһы тарихе) / проф. К. З. Әхмәров; яуаплы ред. Ж. Ғ. Киекбаев.—Өфө: Башҡортостан китап нәшр. , 1972.—132, [2] б.; 21.—Тит. л. парал. рус..— Библиогр.: с. 129-133 и в подстроч. примеч.http://fotki.yandex.ru/users/ievlampiev/album/67013/

7. Examples and Figures

Note: The figure numbers were not changed in doing this revision. As the revision no longer deals with some characters which were not accepted by UTC #127, some figures referring to these letters were dropped. In consequence, the sequence of figure numbers contains gaps.

Fig. 1.Tatar alphabet sorting from [1]

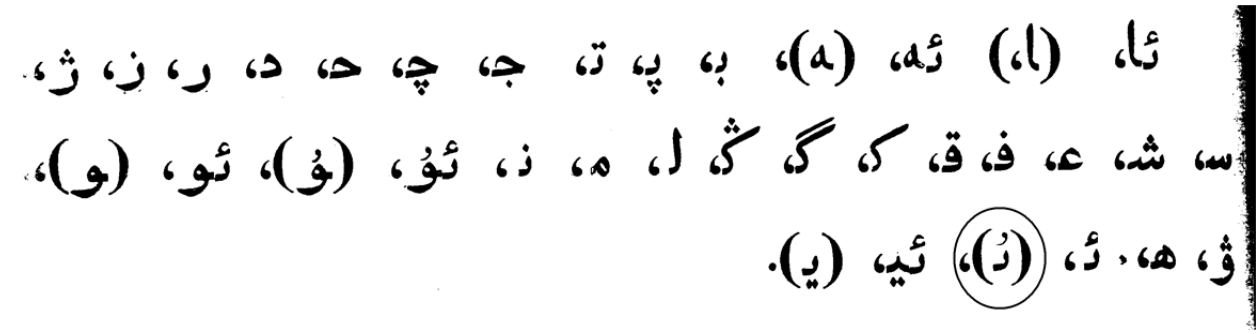


Fig. 3. Usage of words with low alef and without it [1]

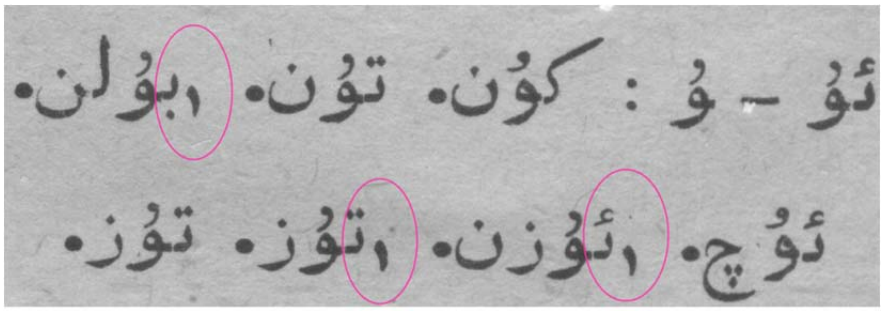


Fig. 6. Simultaneous usage of harakas (pink frames) and modifier dammas over yeh (blue frames). [See book title on http://fotki.yandex.ru/users/ievlampiev/view/141416?page=6]

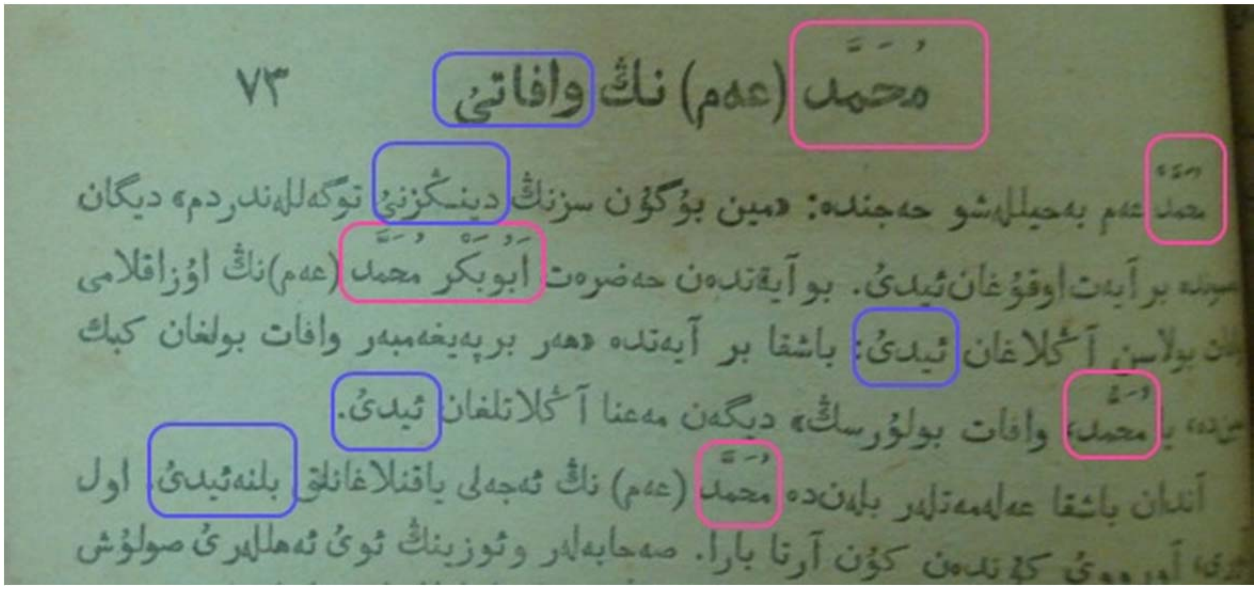
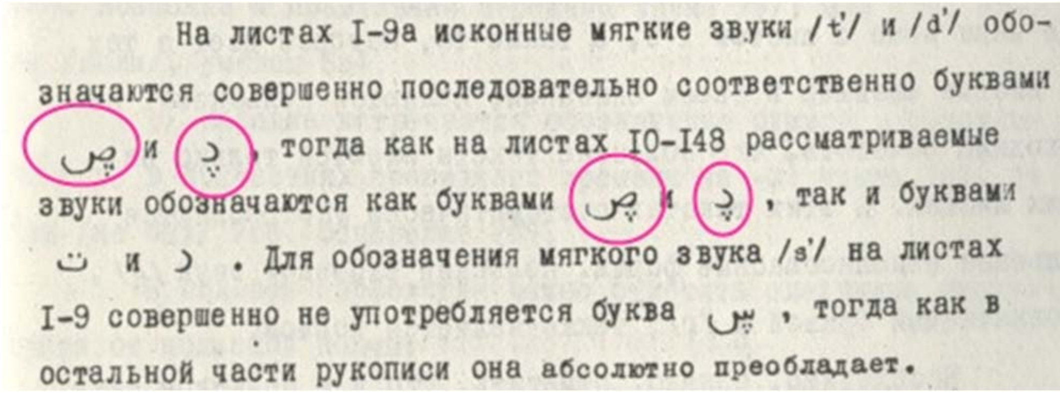


Fig. 7. Usage of Belarusian Arabic. Proposed letters are in frames. [3]



На листах I-9а исконные мягкие звуки /t'/ и /d'/ обозначаются совершенно последовательно соответственно буквами ڝ и ڊ, тогда как на листах I0-I48 рассматриваемые звуки обозначаются как буквами ڝ и ڊ, так и буквами ث и ذ. Для обозначения мягкого звука /s'/ на листах I-9 совершенно не употребляется буква ڛ, тогда как в остальной части рукописи она абсолютно преобладает.

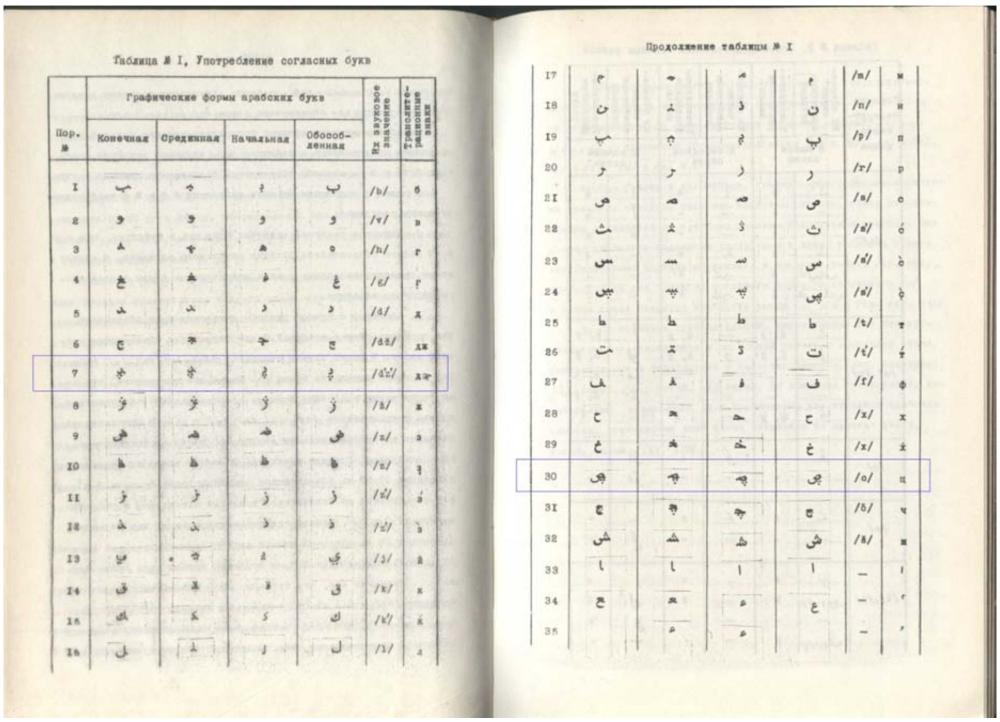
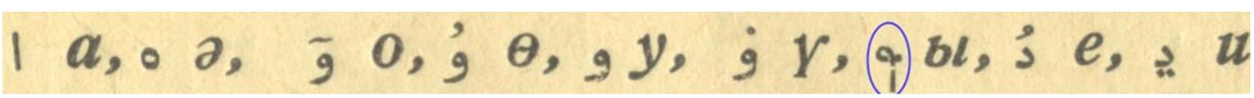


Fig. 8. Usage of straight waw [4]



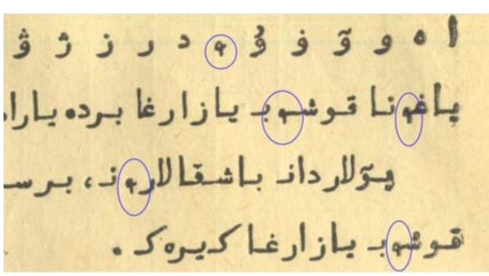


Fig. 9. Usage of low alef (marked red) in the Bashkir language [2]

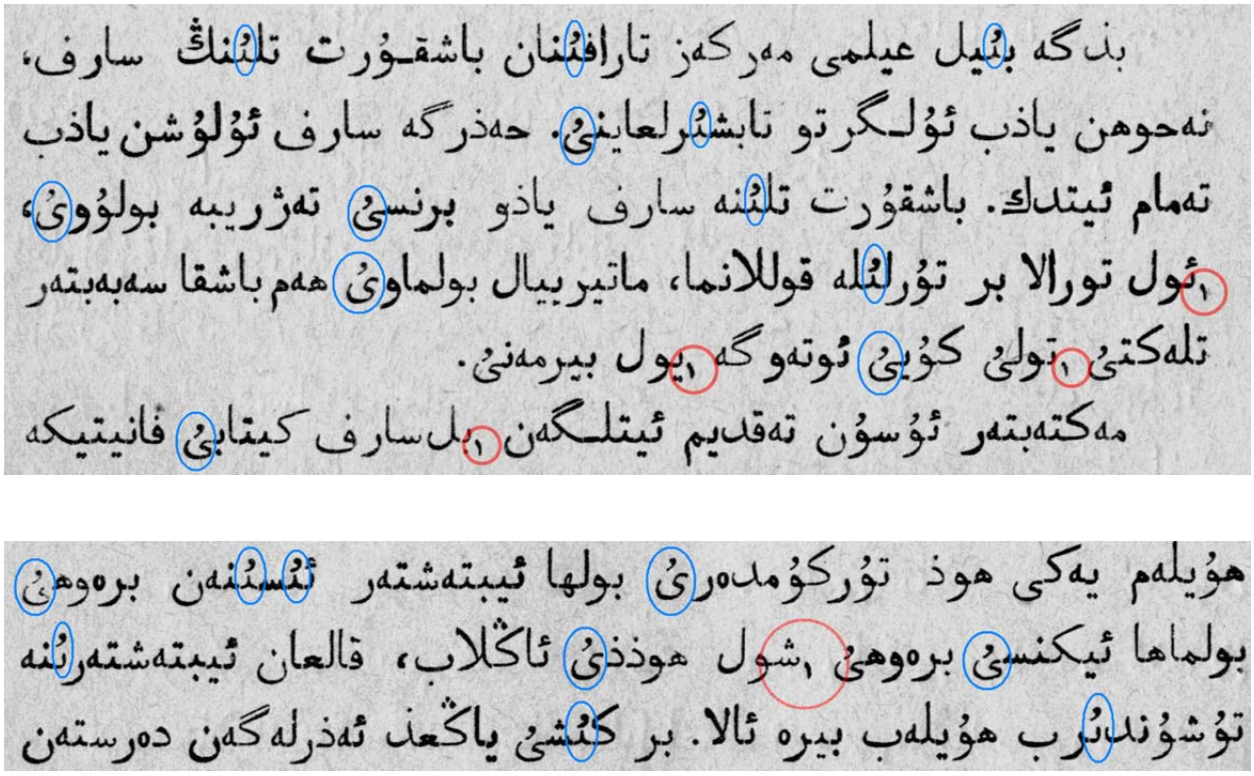


Fig. 10.UsageofproposedlettersinBelarusiantext [Кітабы — унікальная зьява ў беларускай мове/ В. І. Несьцяровіч http://www.pravapis.org/art_kitab1.asp].

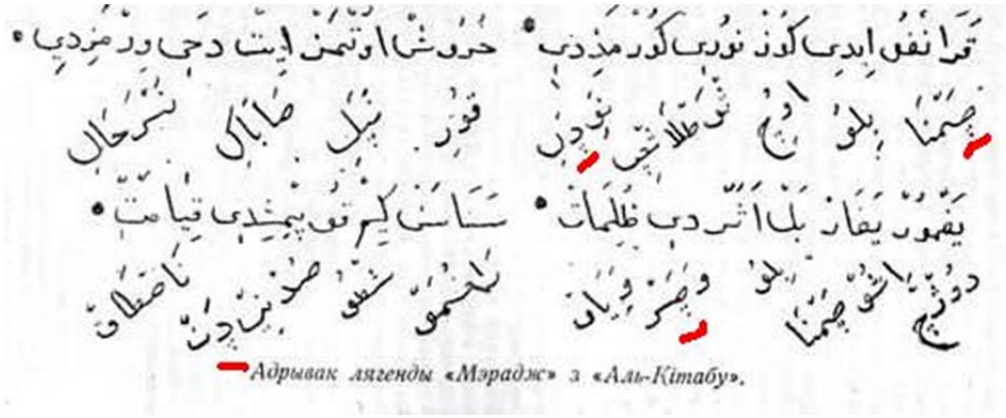
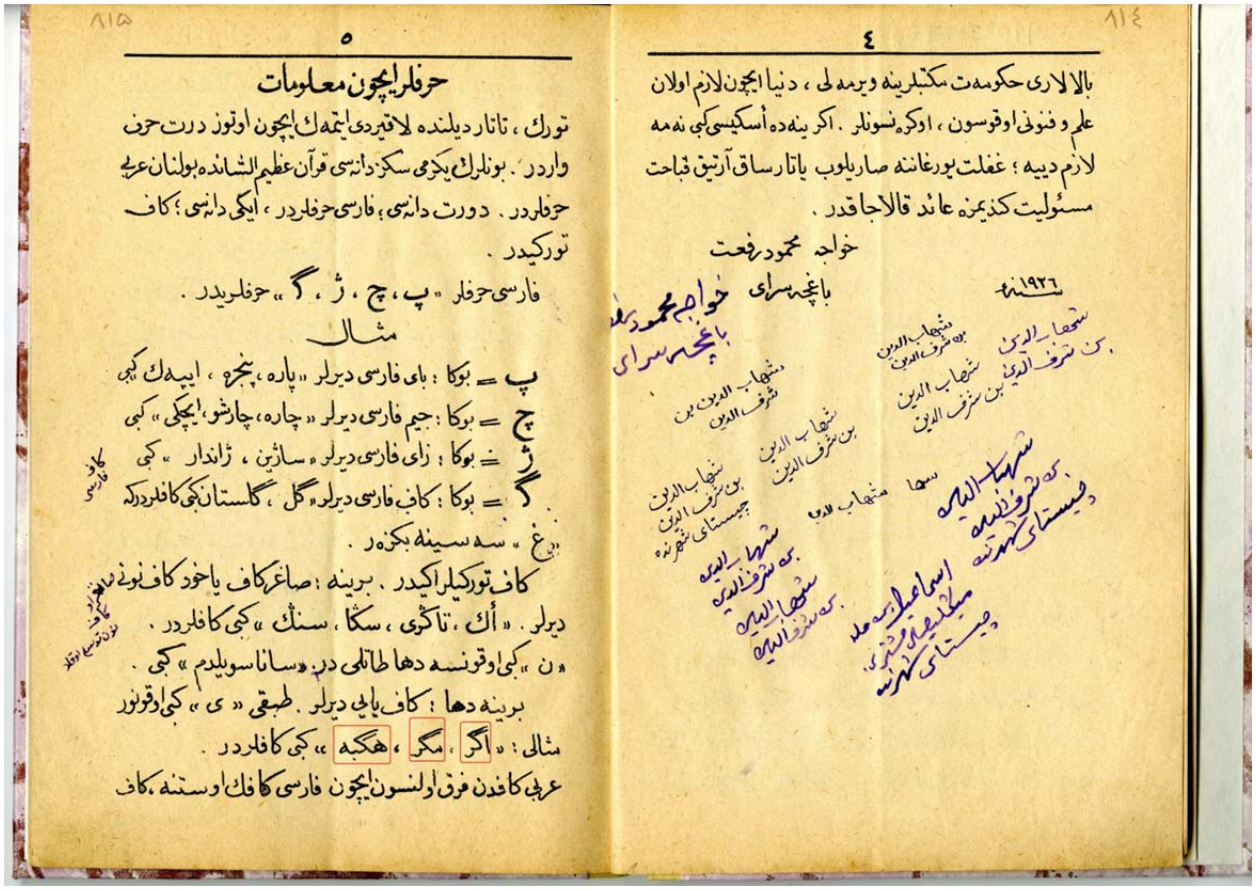


Fig. 14. Page of a Crimean Tatar primer for reading Qur'an [title is here http://fotki.yandex.ru/users/ievlampiev/view/158092?page=0]



٤

بالا لاری حکومه‌ت مکتبلرینه ویرمه لی ، دنیا ایچون لازم اولان علم و فنونی اوقوسون ، اوکره‌نسونلر . اکر بنه ده أسکیسی کبی نه‌مه لازم دییه ؛ غفلت یورغانینه صاریلوب یاتارساق آرتیق قباحت مسئولیت کذیمزه عائد قالاجاقدر .

خواجه محمود رفعت

باغچه‌سرای ۱۹۲۶نجی سنه

٥

حرفلرایچون معلومات

تورك ، تاتار دیلنده لاقیردی ایتمه‌ك ایچون اوتوز درت حرف واردر . بونلرك یکرمی سکزی دانیسی قرآن عظیم الشانده بولنان عربی حرفلردر . دورت دانیسی ، فارسی حرفلردر ، ایکی دانیسی ؛ کاف تورکیدر .

فارسی حرفلر « پ ، چ ، ژ ، گ » حرفلریدر .

مثال

پ = بوکا : بای فارسی دیرلر « پاره ، پنجره ، اپیه‌ك » کبی

چ = بوکا : جیم فارسی دیرلر « چاره ، چارشو ، ایچکی » کبی

ژ = بوکا : زای فارسی دیرلر « ساژین ، ژاندار » کبی

گ = بوکا : کاف فارسی دیرلر « گل ، گلستان » کبی کافلردرکه

غ « سه سینه بکزه‌ر .

کاف تورکیلر اکیدر . برینه : صاغر کاف یاخود کاف نونی دیرلر . « أك ، تاکری ، سکا ، سنك » کبی کافلردر .

« ن » کبی اوقونسه دها طاتلی در « ساناسویلدم » کبی .

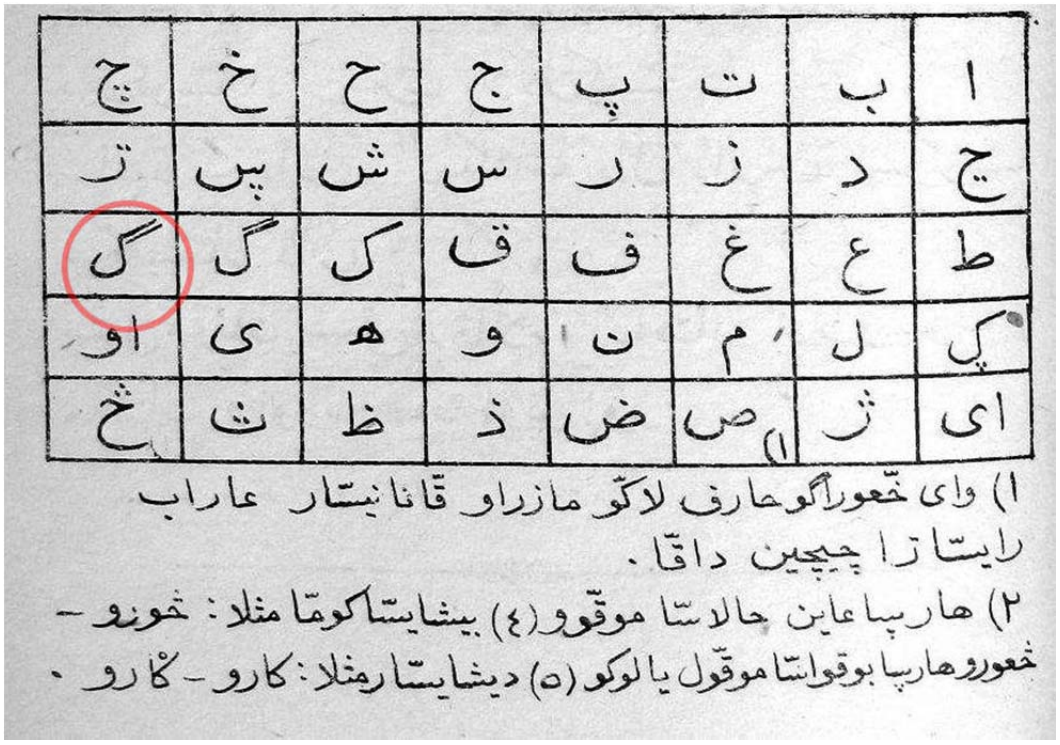
برینه دها : کاف یایی دیرلر . طبقی « ی » کبی اوقونور

مثالی : « اگر ، مگر ، هگبه » کبی کافلردر .

عربی کافدن فرق اولنسون ایچون فارسی کافك اوستنه ، کاف

Fig. 15. Title of Chechen primer and its p. 42 [title is on the image, http://fotki.yandex.ru/users/ievlampiev/album/66437/]

Fig. 16. Lak Arabic alphabet table [Adyghe alphabet table, retrieved from Wikipedia 2010-05-08.
Accompanying information:
Lak arabic alphabet.JPG
English: Lak arabic alphabet from 1925 book
Source: Букварь на лакском языке. Primer auf Lak. Буйнакск, 1925 Buinaksk, 1925]



ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARYFORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html **.**
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** ***Revised proposal to encode Arabic characters used for Bashkir, Belarusian, Crimean Tatar, and Tatar languages***
2. Requester's name: *Ilya Yevlampiev, Karl Pentzlin, Nurlan Joomagueldinov*
3. Requester type (Member body/Liaison/Individual contribution): *Expert Contribution*
4. Submission date: *20 May 2011*
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: *No*

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): *No*
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *Yes*
      Name of the existing block: *Arabic Extended-A*
2. Number of characters in proposal: *5*
3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary		B.1-Specialized (small collection)	X	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?
   Lorna Priest, SIL International
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
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*
   Sorting and linguistic representations are discussed in the proposal
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 http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for

[1] Form number: N3702-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11)

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 examples in proposal*
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *once used by 3 million people*
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *Historically common*
   Reference:
5. Are the proposed characters in current use by the user community? Historical use
   If YES, where? Reference:
6. 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? *Should be placed with similar characters*
   If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?
8. 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?
   If YES, reference:
9. 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? *Discussion in proposal*
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? *No*
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences? *No*
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. 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)
13. Does the proposal contain any Ideographic compatibility character(s)? *No*
    If YES, is the equivalent corresponding unified ideographic character(s) identified?
    If YES, reference: