L2/02-274

Proposal for extensions to the Arabic block

Date: July 16, 2002
Author: Jonathan Kew, SIL International
Address: Horsleys Green
High Wycombe
Bucks HP14 3XL
England
Tel: +44 1494 682306
Email: jonathan_kew@sil.org

A. Administrative

1. Title	Proposal for extensions to the Arabic block
2. Requester's name	SIL International (contacts: Jonathan Kew, Peter Constable)
3. Requester type	Expert contribution
4. Submission date	July 16, 2002
5. Requester's reference	
6a. Completion	This is a complete proposal
6b. More information to be provided?	No

B. Technical — General

1a. New script? Name?	No
1b. Addition of characters to existing block? Name?	Yes — Arabic
2. Number of characters in proposal	4
3. Proposed category	A
4. Proposed level of implementation and rationale	1
5a. Character names included in proposal?	Yes
5b. Character names in accordance with guidelines?	Yes
5c. Character shapes reviewable?	Yes
6a. Who will provide computerized font?	Jonathan Kew, SIL International
6b. Font currently available?	Yes
6c. Font format?	TrueType
7a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?	Yes
7b. Are published examples (such as samples from newspapers, magazines, or other sources) of use of proposed characters attached?	Yes
8. Does the proposal address other aspects of character data processing?	Yes, suggested character properties are included

C. Technical — Justification

1. Has this proposal for addition of character(s) been submitted before?	No
2a. Has contact been made to members of the user community?	Yes
2b. With whom?	During several years in Pakistan, Jonathan Kew worked with Pakistani communities in computerized text editing and publishing.
3. Information on the user community for the proposed characters is included?	Yes
4. The context of use for the proposed characters	Books published in Pathwari, Gojri, Shina, Saraiki, and other languages using Arabic script in Pakistan and India
5. Are the proposed characters in current use by the user community?	Yes
6a. Must the proposed characters be entirely in the BMP?	Yes
6b. Rationale?	Contemporary characters in common use
7. Should the proposed characters be kept together in a contiguous range?	No
8a. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	No
8b. Rationale for inclusion?	N/A
9a. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?	No
9b. Rationale for inclusion?	N/A
10. Does the proposal include use of combining characters and/or use of composite sequences?	No
11. Does the proposal contain characters with any special properties?	All proposed characters have Arabic joining behavior (joining classes are given below)

D. SC2/WG2 Administrative

To be completed by SC2/WG2

1. Relevant SC2/WG2 document numbers	
2. Status (list of meeting number and corresponding action or disposition)	
3. Additional contact to user communities, liaison organizations, etc.	
4. Assigned category and assigned priority/time frame	
Other comments	

I. Proposal

This proposal presents several extensions to the Arabic script that are not currently included in the UCS repertoire. These characters are used in languages of Pakistan and India where Arabic script has been adopted as the basis of the orthography, but sound systems that differ from those of Arabic, Farsi, Urdu, etc., have necessitated the creation of additional characters.

In most cases, the characters proposed here have been used in writing several languages in the South Asia region; examples of the use of each character in one or more languages are included, but these should not be interpreted as representing the entire scope of use. It also seems likely that as additional minority languages in the same areas establish orthographic conventions, they may adopt some of these characters.

1. SEEN with four dots above

Representative glyph	*Suggested name*	*Examples*
ﺵ	ARABIC LETTER SEEN WITH FOUR DOTS ABOVE	See figures 1, 2

Writers of the Shina language in Kashmir, needing to write a retroflex /s/ letter, have adopted the convention of using a letter based on SHEEN but with four dots in place of three.

As a pattern of four dots is not normally seen in the area (though it does occur in southern Pakistani languages such as Sindhi; see U+067F, U+0680, etc.), some writers have tended to replace the four dots with two horizontal lines. This is seen in some of the examples (below). However, as it is common in handwritten script to see horizontal pairs of dots written as a single line, it seems appropriate to consider this a glyph or stylistic variation, encoding the character as SEEN WITH FOUR DOTS and leaving it to font designers to determine whether to offer glyphs with the dots replaced by lines. (It does appear from [Taj89] that some writers specifically choose to write the retroflex /s/ with horizontal lines, even when well-formed dots are used on other letters. Nevertheless, it seems clear that the form with four dots and that with two lines are glyph variants of the same underlying character.)

It may be noted that document [Akbar85] also shows a character with the form of HAH WITH FOUR DOTS ABOVE used by a Shina writer (to represent the phoneme /ts/). The information currently available suggests that this character has not been as widely adopted as the others shown here, with other writers using U+0685 or U+0697 for this sound, so the case for adding it to the Unicode repertoire may be less clear-cut, at least until further data is obtained.

2. Retroflex NOON with dot

Representative glyph	*Suggested name*	*Examples*
ﻧ	ARABIC LETTER NOON WITH DOT AND SMALL TAH	See figures 3, 4, 5, 6, 7

A number of languages use this character to represent a retroflex /n/. The use of a small TAH as part of an Arabic-script letter to indicate retroflexion is derived from Urdu usage, but Urdu itself does not have a retroflex /n/.

Sindhi has this sound, and writes it with a NOON where the dot is replaced by the small TAH (rather than the small TAH being added, while retaining the dot as well). However, this is not an option in languages where the orthography is based on Urdu, as the initial and medial forms would be indistinguishable from the retroflex /t/ (U+0679). Writers of such languages have therefore devised this letter, which is not yet encoded in Unicode.

To reduce the likelihood of confusion with the Sindhi retroflex /n/ letter (encoded as U+06BB ARABIC LETTER RNOON), the suggested name explicitly mentions the presence of the dot in this character; a name such as NOON WITH SMALL TAH might be taken to imply that the small TAH replaces the dot instead of being an addition.

3. Retroflex DAL with two dots below

Representative glyph	*Suggested name*	*Examples*
ڈ	ARABIC LETTER DAL WITH TWO DOTS VERTICALLY BELOW AND SMALL TAH	See figures 3, 4, 5

This character is used by Saraiki writers to represent an implosive retroflex /d/. Other Saraiki retroflexes are written with a small TAH above, following Urdu (e.g., U+0679, U+0688); other implosives are written with two dots vertically below, following Sindhi practice (e.g., U+067B, U+06B3). The natural tendency for a writer needing to write an implosive retroflex /d/, then, is to add the dots below a character like U+0688. Examples can readily be found in Saraiki newspapers and other books.

4. NOON with small V

Representative glyph	*Suggested name*	*Examples*
ڼ	ARABIC LETTER NOON WITH SMALL V	See figures 8, 9

The Gojri community is found in both India and Pakistan, and there has been a significant amount of literature published in both countries from the 1980s onwards. While there has been some variation in orthographic conventions, there is widespread use of a NOON WITH SMALL V to represent the retroflex nasal consonant. (Gojri writers also use a LAM WITH SMALL V, as seen in the examples, but this is already encoded at U+06B5.)

Phonologically, this character is used (at least in the Gojri language) to write the same sound as Sindhi RNOON or the NOON WITH DOT AND SMALL TAH proposed here, but it clearly represents a distinct choice of extended-Arabic character for this sound; the SMALL V and SMALL TAH show two different conventions for the creation of new Arabic-script letters. The fact that the phoneme being written may be the same is irrelevant to the encoding of the written characters.

5. Summary of requested characters

The four extended Arabic characters proposed are summarized below, with their important properties for the Unicode character database. It is proposed that these characters are added at some of the as-yet-unassigned locations in the early columns of the Arabic block, as there are no remaining positions in the "extended" part of the U+06xx block. This means that some "extended" characters will be interspersed with the "standard" Arabic letters; while this seems a little untidy, no better option appears to be available.

Representative glyph	*Suggested USV*	*Suggested character name*	*General category*	*Combining class*	*Bidi category*	*Shaping group*
ڈٟ	U+063C	ARABIC LETTER DAL WITH TWO DOTS VERTICALLY BELOW AND SMALL TAH	Lo	0	AL	DAL
ۺ	U+063D	ARABIC LETTER SEEN WITH FOUR DOTS ABOVE	Lo	0	AL	SEEN
ںٟ	U+063E	ARABIC LETTER NOON WITH DOT AND SMALL TAH	Lo	0	AL	NOON
ںٮ	U+063F	ARABIC LETTER NOON WITH SMALL V	Lo	0	AL	NOON

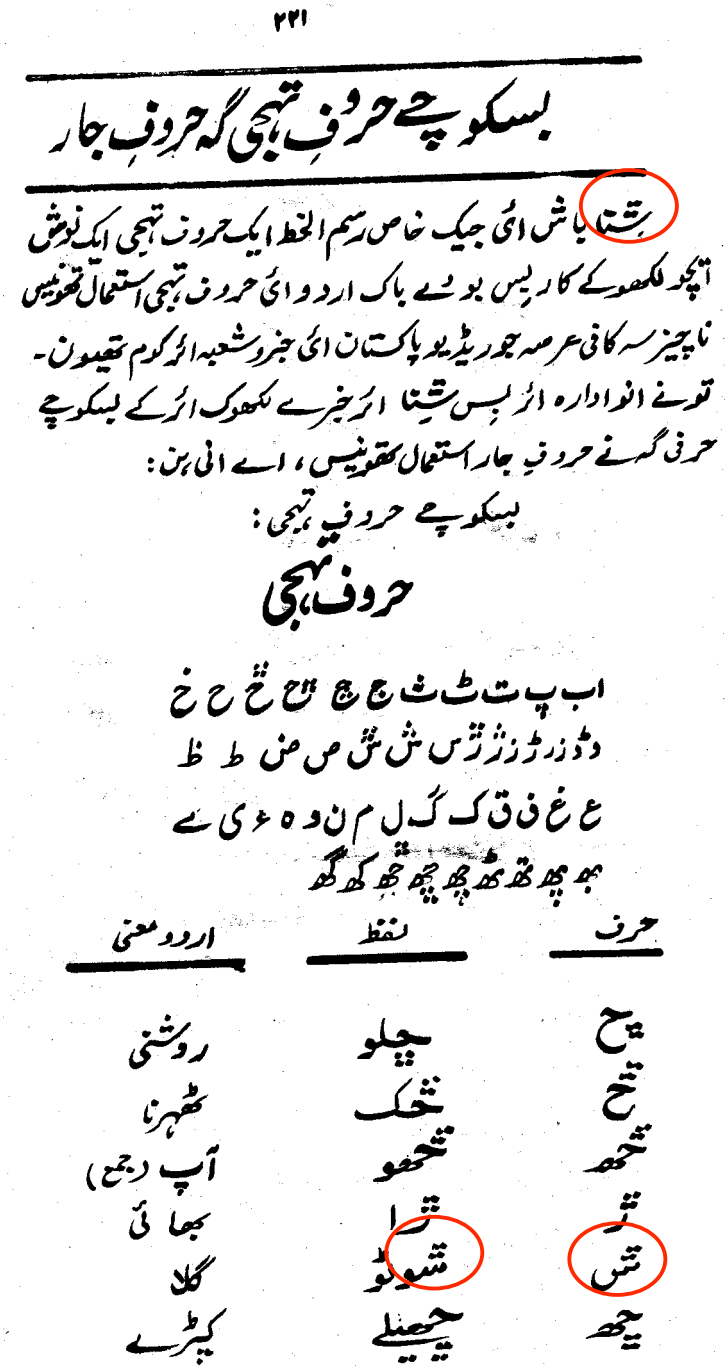
The shaping behavior of these characters is in accordance with the shaping groups indicated above; the following table illustrates how the characters would be shaped in a basic Arabic type design.

Name	*Isolate*	*Initial*	*Medial*	*Final*
ARABIC LETTER DAL WITH TWO DOTS VERTICALLY BELOW AND SMALL TAH	ڈٟ			ڈٟ
ARABIC LETTER SEEN WITH FOUR DOTS ABOVE	ۺ	ۺـ	ـۺـ	ـۺ
ARABIC LETTER NOON WITH DOT AND SMALL TAH	ںٟ	ںٟـ	ـںٟـ	ـںٟ
ARABIC LETTER NOON WITH SMALL V	ںٮ	ںٮـ	ـںٮـ	ـںٮ

II. Examples of usage

The examples shown here are drawn largely from books and newspapers published in Pakistan (an Indian example is also included), written in Shina, Saraiki, Pathwari, and Gojri. However, it can be expected that as literacy becomes more widespread among minority language communities in South Asia, many of these writing conventions may be "borrowed" by neighboring communities. The languages cited here serve to demonstrate the need to encode these characters, but should not be assumed to be the only users of them.

1. Shina



۲۲۱

بسکو چے حروفِ تہجی گہ حروفِ جار

شِتا باش ائی جیک خاص رسم الخط ایک حروف تہجی ایک نوش آچکو لکھوکے کا ریس بوے باک اردو ائی حروف تہجی استعمال تھونیس ناچیز سہ کافی عرصہ جو ریڈیو پاکستان ائی جنرو شعبہ ائرکرم تھیون۔ تونے انو ادارہ ائر بس شِتا ائر خبرے لکھوک ائر کے بسکو چے حرفی گہ نے حروفِ جار استعمال تھونیس، اے انی ہن:

بسکو چے حروفِ تہجی:

حروفِ تہجی

ا ب پ ت ٹ ث ج چ ح خ
د ڈ ذ ر ڑ ز ژ س ش ص ض ط ظ
ع غ ف ق ک گ ل م ن و ہ ء ی ے
بھ پھ تھ ٹھ جھ چھ ٹھ کھ گھ

حرف	لفظ	اردو معنی
چ	چلو	روشنی
خ	خک	ٹھہرنا
خھ	خھو	آپ (جمع)
ژ	ژا	بھائی
ش	شوٹو	گلا
چھ	چھیلے	کپڑے

Figure 1: [Akbar85], page 221

Figure 2: [Taj89], page 29: examples of SEEN WITH FOUR DOTS ABOVE written using two lines in place of four dots (a practice also used for other four-dot letters in this book)

2. Saraiki

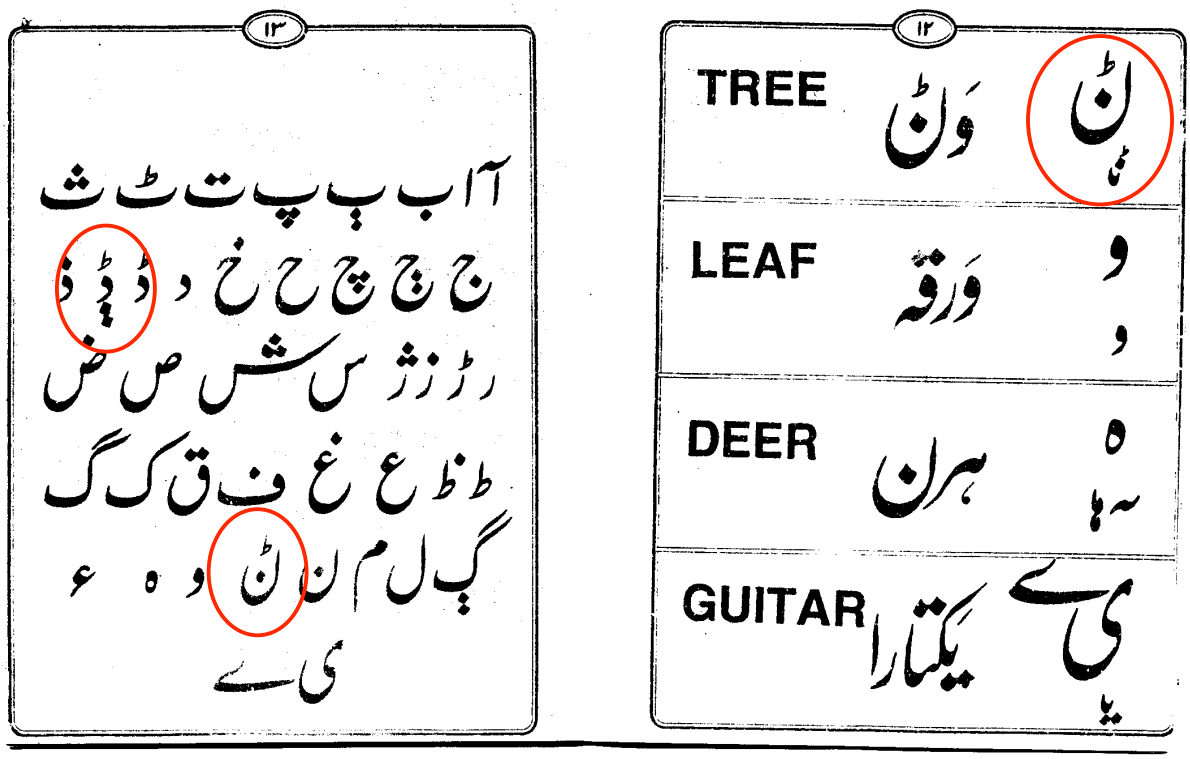


Figure 3: [Mughal94], pages 12–13

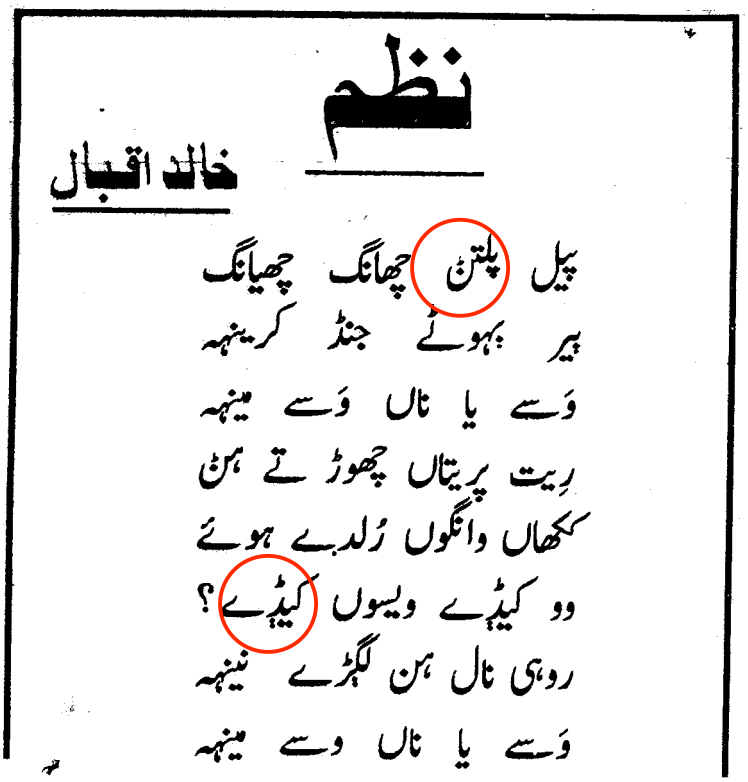


Figure 4: Saraiki poem printed in [Jhoke], 29 January 2002. The small size of the added TAH and TWO DOTS on the special Saraiki letters arises because these letters are not supported in Urdu software used to produce this newspaper, so these have been added manually. A similar size anomaly can be seen in the dots under BEEH (U+067B) in line 2 of the poem; this character is also not used in Urdu, so the Urdu BEH (with one dot) has been typed, and a second dot added manually.

صرف

صرف لغت وچ بدلݨ یا ہیر پھیر کرݨ کوں آہدن ۔ علم وچ حروف تے انھاں دی حرکات (زبر ، زیر پیش) دے ادل بدل نال انجو انج لفظ تے انھاں دے وکھرے وکھرے معنی حل آندن ۔ ایں گالھوں ایندا ناں علم صرف رکھیا گئے ۔

نحو

او علم ہے جیندے وچ کلام یعنی گالھ دے مختلف حصیں کوں صحیح طرحاں جوڑݨ تے ولݨ آندے تے انھاں حصیاں دے آپس دے تعلق دا پتہ لگدے ۔ تے انھاں غلطیاں کنوں بچیندے ۔ جنہاں دی وجہ نال کوئی گالھ سمجھ نہ آوے ، یا اوندا مطلب و گڑ ونجے ۔

حروف تہجی

کہیں زبان دیاں سادیاں اوازاں کوں حروف تہجی آکھیا ویندے ۔ جیویں جو ۔ ب ۔ پ ۔ ک ۔ م ۔ ل وغیرہ

سرائیکی دے مخصوص اضافی حروف تہجی

اردو دے سارے دے سارے حروف تہجی سرائیکی وچ وی استعمال کیتے ویندن البتہ خاص خاص اوازاں کیتے سرائیکی وچ پنج حروف اضافی طور تے شامل کیتے گئن ۔ او ایہ ہن ۔

ب ------ بلی ۔ بال ج ------ جالی ۔ جنج

ڈ ------ ڈانگ ۔ ڈلی گ ------ گاں ۔ لنگ

ٹ ------ وٹ ۔ پانی

Figure 5: [Bhaya84], introductory page explaining special Saraiki letters

3. Pathwari/Pahari

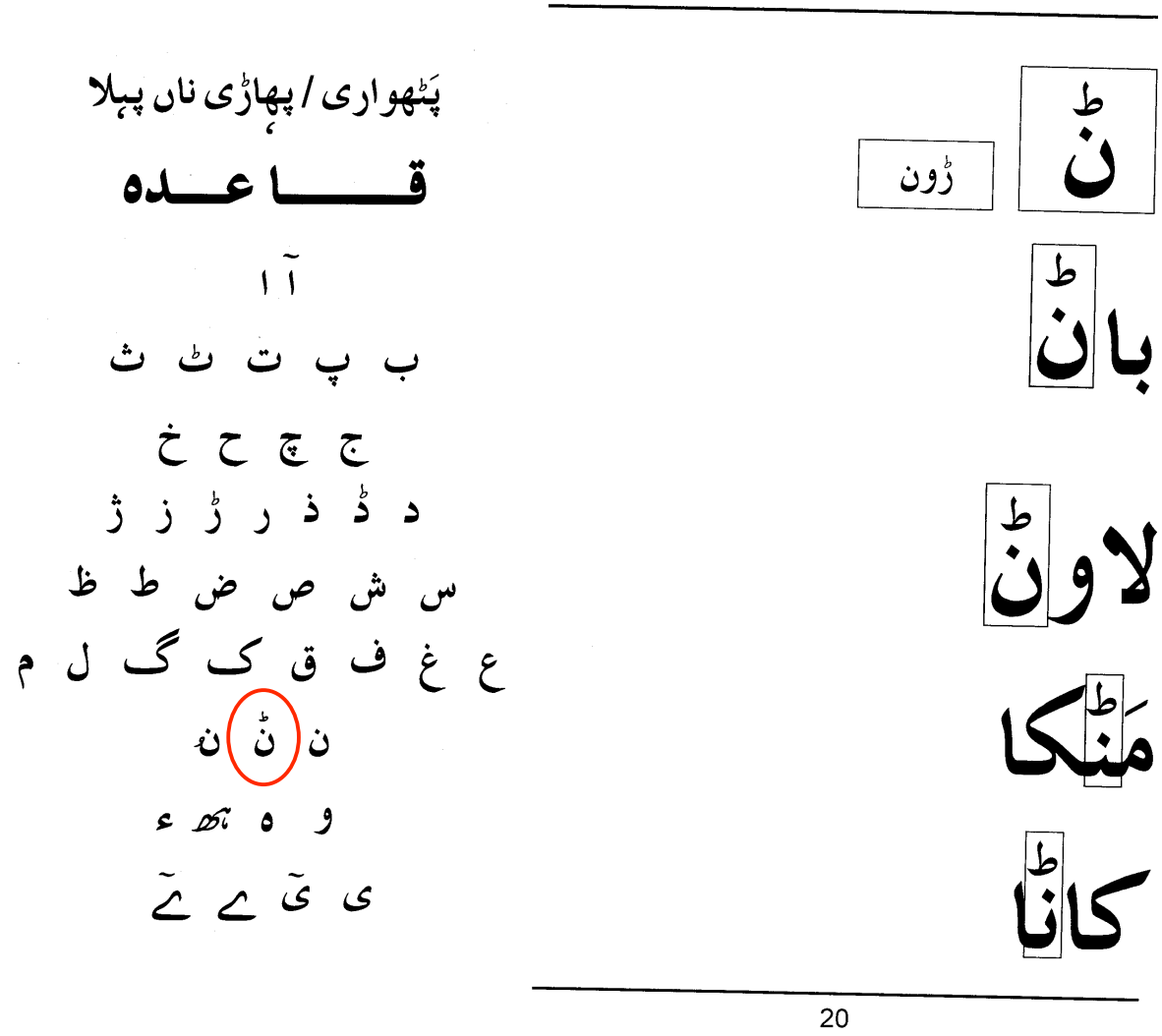


Figure 6: [Chitka], inside front cover and page 20. The alphabet chart also shows a second "modified NOON" character, but it is unclear whether this form can be considered well enough established for standardization.

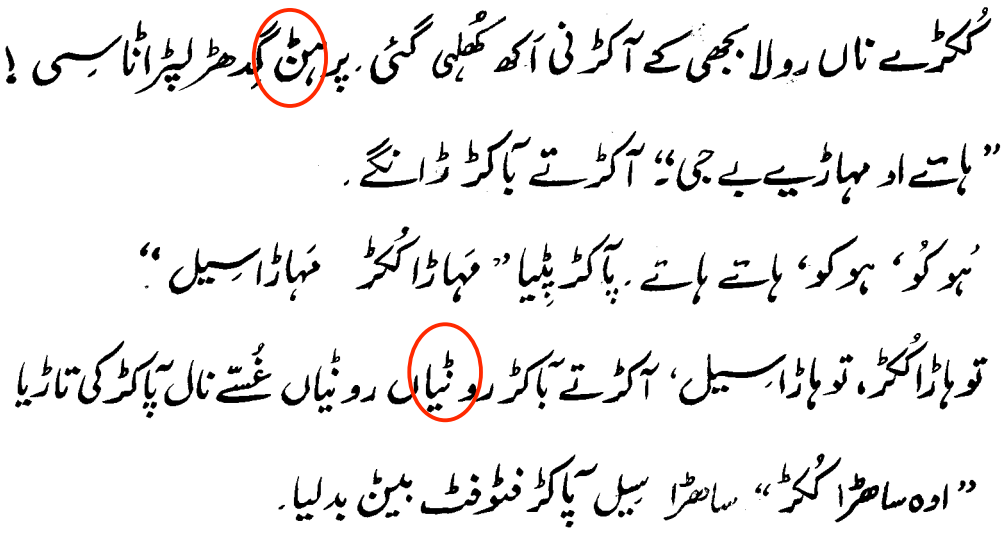
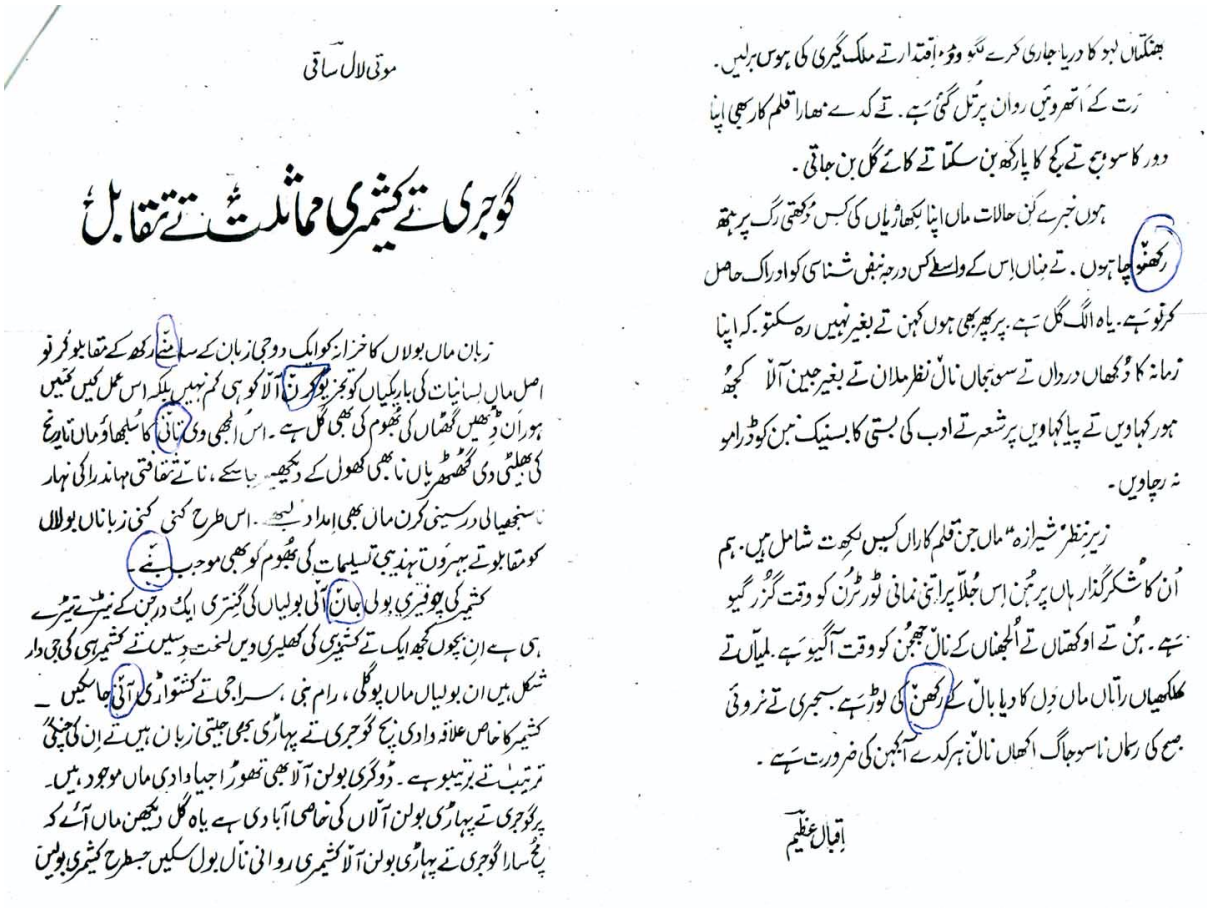


Figure 7: [Mehmood01], extract from page 41.

4. Gojri



موتی لال ساقی

گوجری تے کشمیری مماثلت تے تقابل

زبان ماں بولاں کا خزانہ کو ایک دوجی زبان کے سامنے رکھ کے مقابلو کرنو اصل ماں لسانیات کی بارکیاں کو تجزیو کرن آلا کو ہی کم نہیں بلکہ اس عمل کیں تمیں ہور اَن دِٹھیں گھماں کی بھوم کی بھی گل ہے۔ اس الجھی وی تانی کا سُلجھاؤ ماں تاریخ کی بھلیّ دی گھُٹھڑیاں نا بھی کھول کے دیکھیو جا سکے، نا تے ثقافتی ہماندرا کی نہار سنجھیالی درسینی کرن ماں بھی امداد لیجے۔ اس طرح کنی کنی زباناں بولاں کو مقابلو تے ہیروں تہذیبی تسلیمات کی بھوم کو بھی موجب بنے۔

کشیر کی چوفیری بولی جان آلی بولیاں کی گنتری ایک درجن کے نیڑے تیڑے ہی ہے اِن بچوں کجھ ایک تے کشیری کی کھلیری ویں لغت دِسیں تے کشیر ہی کی جی دار شکل ہیں ان بولیاں ماں پوگلی، رام بنی، سراجی تے کشتواڑی آنی جا سکیں۔ کشیر کا خاص علاقو وادی پچ گوجری تے پہاڑی بھی جیتی زبان ہیں تے اِن کی اپنی ترتیب تے ترتیبو ہے۔ ڈوگری بولن آلا بھی تھوڑا جیا وادی ماں موجود ہیں۔ پر گوجری تے پہاڑی بولن آلاں کی خاصی آبادی ہے یاہ گل دیکھن ماں آئے کہ پچ سارا گوجری تے پہاڑی بولن آلا کشیری روانی نال بول سکیں جسطرح کشیری بولیں

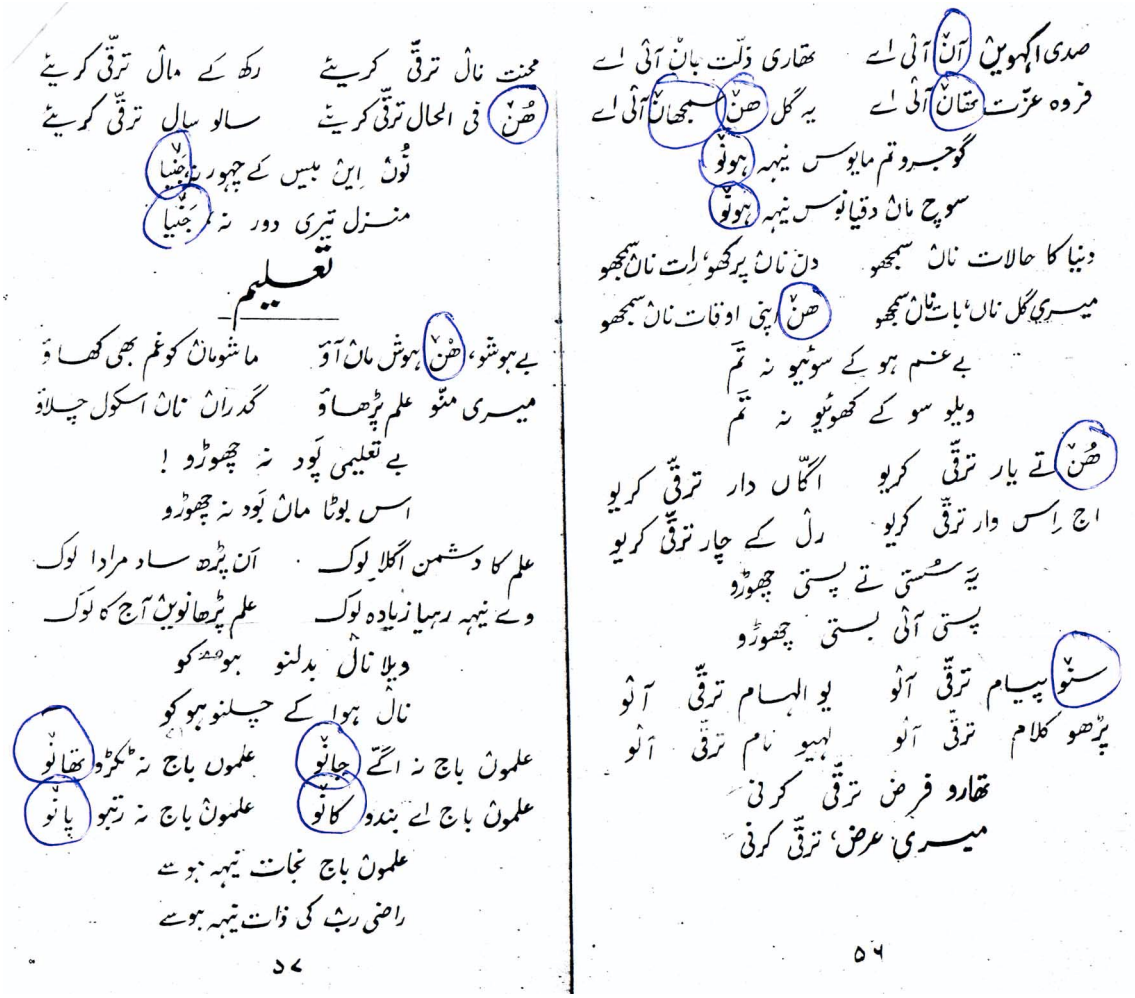
بھٹکاں لہو کا دریا جاری کرے لگو وڈو اقتدار تے ملک گیری کی ہوس برلیں۔
رت کے اتھرو ئیں روان پرتل گئی ہے۔ تے کدے نھارا قلم کار بھی اپنا
دور کا سوچ تے بچ کا پارکھ بن سکتا تے کائے گل بن جاتی۔

ہوں خبرے کن حالات ماں اپنا بکھاریاں کی کسک مرکھتی رگ پر ہتھ
رکھنو چاہتوں۔ تے مناں اس کے واسطے کس درجہ نبض شناسی کو ادراک حاصل
کرنو ہے۔ یاہ الگ گل ہے۔ پر پھر بھی ہوں کہن تے بغیر نہیں رہ سکتو۔ کہ اپنا
زمانہ کا دُکھاں درداں تے سوچاں نالّ نظر ملان تے بغیر جیون آلا کجھ
ہور کہا دیں تے پیا کہاویں پر شعر تے ادب کی بستی کا بسنیک نہیں کوڑ رامو
نہ رچا دیں۔

زیرِ نظر "شیرازہ" ماں جن قلم کاراں کیس لکھت شامل ہیں۔ ہم
اُن کا شکر گذار ہاں پر جن اِس جلاّ پرا اتنی نمانی ٹور ٹرن کو وقت گزر گیو
ہے۔ ہُن تے او کھاں تے اُلجھاں کے نالّ جھجھن کو وقت آگیو ہے۔ لمیاں تے
کھلکھیاں راتاں ماں دِل کا دیا بالّ کے رکھن کی لوڑ ہے۔ سجری تے نروئی
صبح کی رساں نا سوجاگ اکھاں نالّ نہار کے آنکھن کی ضرورت ہے۔

اقبال عظیم

Figure 8: [Azeem96], unnumbered pages



صدی اکھوین آنْ آئی اے تھاری ذلّت جانْ آئی اے
فروہ عزّت تھانْ آئی اے یہ گل ھنْ سمجھانْ آئی اے
گو جبرو تم مایوس نیہہ ہونو
سوچ مانْ دقیانوس نیہہ ہونو
دنیا کا حالات نانْ سمجھو دن نانْ پرکھو رات نانْ سمجھو
میری گل ناں، بات مانْ سمجھو ھنْ اپنی اوقات نانْ سمجھو
بے غم ہو کے سوئیو نہ تم
ویلو سو کے کھوئیو نہ تم
ھنْ تے یار ترقّی کریو اگّاں دار ترقّی کریو
اج اِس وار ترقّی کریو رلْ کے چار ترقّی کریو
یہ سُستی تے پستی چھوڑو
پستی آئی بستی چھوڑو
سنو پیام ترقّی آئو یو الہام ترقّی آئو
پڑھو کلام ترقّی آئو لہیو نام ترقّی آئو
تھارو فرض ترقّی کرنی
میری عرض، ترقّی کرنی

٥٦

محنت نالْ ترقّی کریئے رکھ کے مالْ ترقّی کریئے
ھُنْ فی الحال ترقّی کریئے سالو سال ترقّی کریئے
تُونْ اِیْن ہیں کے چھوریں جنیا
منزل تیری دور نہ رکھیں جنیا

تعلیم

بے ہوشو، ھُنْ ہوش مانْ آؤ ماشومانْ کو غم بھی کھاؤ
میری منّو علم پڑھاؤ گدرانْ نانْ اسکول چلاؤ
بے تعلیمی پَود نہ چھوڑو !
اِس بوٹا مانْ پَود نہ چھوڑو
علم کا دشمن اگلا لوک آن پڑھ ساد مرادا لوک
وے نیہہ رہیا زیادہ لوک علم پڑھانویں آج کا لوک
ویلا نالْ بدلنو ہوسی کو
نالْ ہوا کے چلنو ہو کو
علموں باج نہ اگّے جانْو علموں باج نہ ٹکڑو تھانْو
علموں باج لے بندو کانْو علموں باج نہ رتبو پانْو
علموں باج نجات نیہہ ہوسے
راضی رب کی ذات نیہہ ہوسے

٥٧

Figure 9: [Afaqi], pages 56–57

III. References

[Afaqi] Afaqi, Dr. Sabir. Unknown date. Pegham-in-qalab [Gojri poetry]. Pakistan.

[Akbar85] Akbar, Akbar Hussain. 1985. Sumulo Rasuul (Holy Prophet). Islamabad, Pakistan: Modern Book Depot.

[Azeem96] Azeem, Iqbal Chaudhry, ed. 1996. Sheeraza [Gojri periodical]. Srinagar: Jammu & Kashmir Academy of Art, Culture and Languages.

[Bhaya84] Bhaya, Bashir Ahmad. 1984, revised 1998. Saraiki quaid te zubandani (Saraiki, father of languages). Bahawalpur, Pakistan: Saraiki Adabi Majlis.

[Chitka] Chitka Committee. Unknown date. Pathwari/Pahari nã pahala qaidah (First Pathwari/Pahari primer). Jhelum, Pakistan: Praala Publishers.

[Jhoke] Daily Jhoke [Saraiki daily newspaper]. Multan, Pakistan.

[Mehmood01] Mehmood, Tariq. 2001. Sayana gidhar (Clever jackal). Jhelum, Pakistan: Praala Publishers.

[Mughal94] Mughal, Shaukat. Saraiki qaidah (Saraiki primer). Multan, Pakistan: Saraiki Isha'ati Idarah.

[Taj89] Taj, Abdul Khaliq. 1989. Shina qaidah (Shina primer). Gilgit, Pakistan: Muhammad Book Stall.