

Proposed Changes to Gurmukhi 3

Document Number: L2/05-344

Submitter's Name: Sukhjinder Sidhu (Punjabi Computing Resource Centre)

Submission Date: 27 October 2005

Abstract

This document addresses issues raised by the Unicode Technical Committee and builds on information in "Proposed Changes to Gurmukhi" (L2/05-088) and "Proposed Changes to Gurmukhi 2" (L2/05-167). Any relevant information from the previous proposals is included within for completeness.

Thanks to everyone who has contributed both time and effort to the research involved in making this document. Special thanks to:

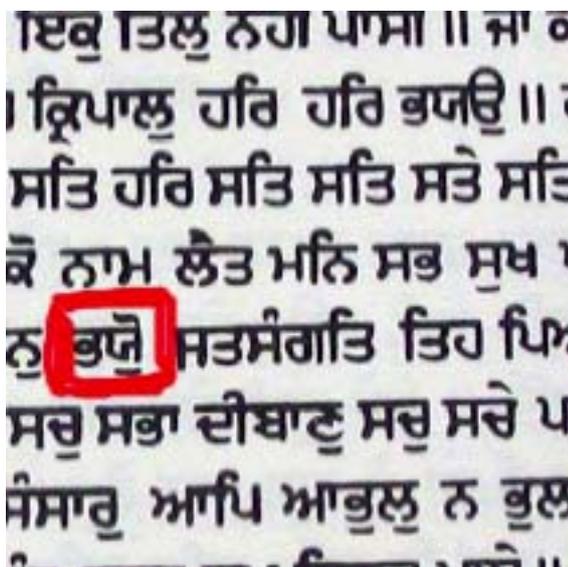
*Jeevan Deol
Manudeep Singh
Serjinder Singh
Kulbir Thind*

...and all the members of the Indic mailing list who have constructively discussed and debated the contents of these proposals.

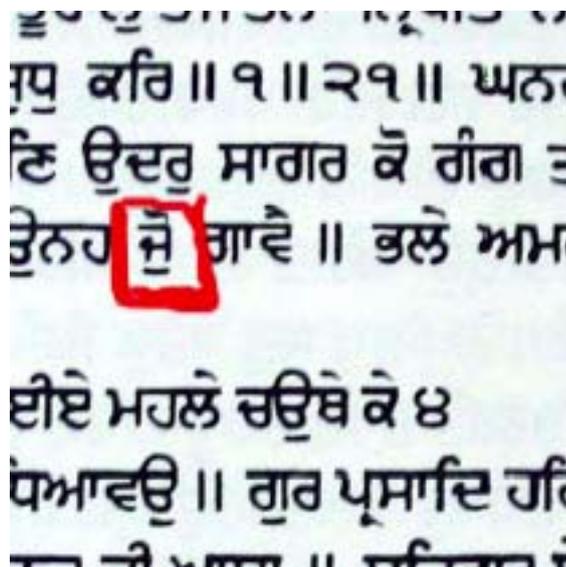
A. Double Vowel Signs

Older Gurmukhi (for example, in the Sikh holy book – the Sri Guru Granth Sahib) is known to use two vowel signs on one consonant. This behaviour is predominantly restricted to Hora (Vowel Sign OO, ੋ, U+0A4B) and Aunkar (Vowel Sign U, ੁ, U+0A41) and their independent form ੳ. This particular combination represents “the metrical shortening of ‘ō’ or lengthening of ‘u’ depending on context.”¹

The additional vowel sign is added to a syllable and lengthens or shortens the vowel based on the original vowel sign. It is designed to keep the meaning of the original word in tact, while indicating how the vowel should be pronounced in poetry.²



A1 SGGS page 1386 (੧੩੮੬)



A2 SGGS page 1396 (੧੩੯੬)

Example

Umāhā (ਉਮਾਹਾ) becomes Ūmāhā (ਊਮਾਹਾ)

Gōbind (ਗੋਬਿੰਦ) becomes Gobind (ਗੋਬਿੰਦ)

Both examples maintain the original meaning of the word while altering the pronunciation.

Proposed Changes

Although this is by far the most common example, other forms do exist, such as:

ਗ੍ਰਿਾਨ

Which shows two vowel signs (i and ā respectively) attached to the conjunct “Gh” and is mentioned on page 1302 of the Guru Granth Sahib.

At present most rendering engines only allow the addition of one vowel sign to a consonant and force any additional vowels to stand alone. We recommend that the Unicode Standard specifically states that more than one vowel sign may attach to a consonant and that a recommended ordering is introduced.

This recommended ordering should be introduced because allowing more than one vowel sign can be a security concern because signs may overlap causing two different text sequences to render identically. For example:

ਤੈ = ਤ + ੋ + ੈ

ਤੈ = ਤ + ੈ

¹ Jeevan Deol, Research Fellow in Indian History, St. John's College, University of Cambridge

² Sahib Singh, *Gurbani Vyakaran (Gurbani Grammar)*, (1994, In Punjabi), p. 405.

In addition, identical visual representations of multiple vowel signs can have different underlying encodings.

In the case of Gurmukhi there are three 'logical' orders for vowel positioning. Using the canonical combining class to assign ordering patterns will be inappropriate for Gurmukhi because all signs have the canonical combining class of 0. There should be a mechanism to prevent more than one vowel sign occupying one side of the consonant.



These three orders are the most suitable for the Gurmukhi script which reads from left-to-right and top-to-bottom:

- Left, Right, Top, Bottom
- Left, Top, Bottom, Right
- Left, Top, Right, Bottom

The following example shows the worst case scenario – when four vowel signs are used all around a consonant. There are three sequences that could be used depending on which of the three vowel ordering methods are employed:

ਬਿੰੀ = ਬ + ਿ + ੀ + ੈ + ੁ OR ਬ + ਿ + ੈ + ੁ + ੀ OR ਬ + ਿ + ੈ + ੀ + ੁ

The following example is identical no matter which vowel ordering method is used:

ਪੁੰ = ਪ + ੋ + ੁ

The final examples show how more than one vowel sign on the same side of the consonant should be made to stand separately:

ਰੇੈ = ਰ + ੇ + ੈ

ਰੇੇੇ = ਰ + ੇ + ੇ + ੇ

This principle should **not** be universally extended to the independent vowels or the standalone forms of Ura (U+0A73), Aira (U+0A05) and Iri (U+0A72). Instead a case-by-case analysis should be considered. It is recommended that the only vowel sign attachment to an independent vowel should be for ੳ + ੁ. The independent form (ੳ) must only be rendered when entering Letter OO (U+0A13) and Vowel Sign U (U+0A41).

Allowing any arbitrary vowel sign to attach to independent vowels can cause security issues in terms of multiple representations for the same visual appearance.

B. Recommended Character Sequences

After the submission of the initial proposals, it became apparent that there were problems with multiple ways of representing Gurmukhi syllables that could not be addressed with normalisation and could potentially pose major security concerns. In response to this, the following rules have been formulated:

- No additional vowel signs should attach to independent vowels – this is especially true for Aira (Letter A) except for pre-defined exceptions.
- Ura and Iri are only designed for singular representation and have no inherent meaning on their own. They should not combine with any signs in the Gurmukhi block.
- Vowel signs must only attach to consonants as indicated in section A.

In response to the recommendations by the UTC, the following table lists the acceptable and unacceptable forms for a given graphical appearance.

Graphical Appearance	Acceptable	Unacceptable
ਆ	U+0A06	U+0A05, U+0A3E
ਇ	U+0A07	U+0A72, U+0A3F
ਈ	U+0A08	U+0A72, U+0A40
ਉ	U+0A09	U+0A73, U+0A41
ਊ	U+0A0A	U+0A73, U+0A42
ਏ	U+0A0F	U+0A72, U+0A47
ਐ	U+0A10	U+0A05, U+0A48
ਓ	U+0A13	U+0A73, U+0A4B
ਔ	U+0A14	U+0A05, U+0A4C
Vowel signs should not be attached to the standalone forms of the vowel bearers (U+0A05, U+0A72 and U+0A73). The pre-composed code points should be used instead.		
ੌ	U+0A4B, U+0A41	U+0A41, U+0A4B
ਊ	U+0A13, U+0A41	U+0A73, U+0A4B, U+0A41 U+0A73, U+0A41, U+0A4B

C. Conjoined Consonants

The following conjuncts should be mentioned in the Unicode standard along with the existing four (Ra, Ha, Va, Ya). Because the number of conjuncts for Gurmukhi is limited, it is requested that they be listed in chapter 9 of the Unicode Standard. Evidence for these forms is provided in the appendix.

Virtually all of the subjoined consonants are equivalent to their full form but without the top bar.

Virama (U+0A4D) + Ga (U+0A17) = ੱ + ਗ = ੱਗ = GURMUKHI PAIRIN GA

Virama (U+0A4D) + Ca (U+0A1A) = ੱ + ਚ = ੱਚ = GURMUKHI PAIRIN CA

Virama (U+0A4D) + Tta (U+0A1F) = ੱ + ਟ = ੱਟ = GURMUKHI PAIRIN TTA

Virama (U+0A4D) + Ttha (U+0A20) = ੱ + ਠ = ੱਠ = GURMUKHI PAIRIN TTHA

Virama (U+0A4D) + Ta (U+0A24) = ੱ + ਤ = ੱਤ = GURMUKHI PAIRIN TA

Virama (U+0A4D) + Da (U+0A26) = ੱ + ਦ = ੱਦ = GURMUKHI PAIRIN DA

Virama (U+0A4D) + Na (U+0A28) = ੱ + ਨ = ੱਨ = GURMUKHI PAIRIN NA

Virama (U+0A4D) + Tha (U+0A25) = ੱ + ਥ = ੱਥ = GURMUKHI PAIRIN THA

Conjuncts may have both post-base and subjoined forms. This is particularly true in the case of Tha and Ya. **As such, font technology and rendering engines should be aware that either form can occur for any conjunct.**

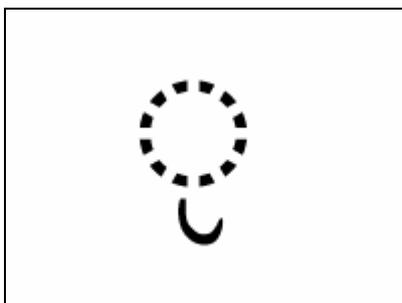
Virama (U+0A4D) + Tha (U+0A25) = ੱ + ਥ = ੱਥ = GURMUKHI ADHA THA³

Virama (U+0A4D) + Ma (U+0A2E) = ੱ + ਮ = ੱਮ = GURMUKHI ADHA MA⁴

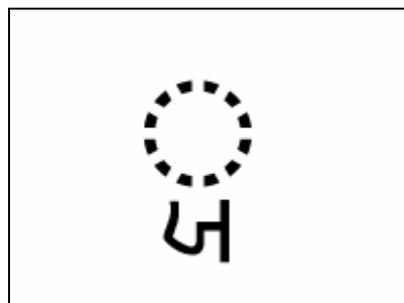
In the case of Ya, there are at least three different conjunct forms. The current one, which is recognised in the Unicode standard, is the modern post-base form. In addition to this, two different subjoined forms are also in use which should be mentioned in the Unicode Standard:



C1 Post-base form of Ya (Modern)



C2 Subjoined form of Ya (Old, SGGS)



C3 Subjoined form of Ya (Old, Other)

³ Linguistic Survey of India Volume IX, 1916, p.627.

⁴ Linguistic Survey of India Volume IX, 1916, p.627.

D. Udaat (ਉਦਾਤ)

Initially it was determined that Udaat was a variant form of subjoined Ha (Pairīn Haha), however after further research this is now believed to be incorrect. This also explains why both subjoined Ha and Udaat are used concurrently in the same document.

Udaat⁵ looks like the Halant or Virama character in Devanagari, but it is not that character. It is found in the Sri Guru Granth Sahib 1188 times⁶. The Udaat is/was used for a non-segmental phoneme (*akhāndi tōni*) known as the high tone⁷. This sign is related to Ha, because Ha itself is used to distinguish tones, but it is not a variant form. Udaat is related to Devanagari Udatta (U+0951) which also indicates a high tone in Sanskrit literature.

High tone is still present in modern Punjabi, however, the Udaat is not used in modern Gurmukhi. In modern Gurmukhi, there are no symbols that highlight the high or low tones. But at places where the Udaat was used earlier, now another symbol known as the Pairīn Haha is being used. This does not mean that the Udaat is equivalent to Pairin Haha. The orthographical rules of Gurmukhi suggest that Pairīn Haha is used for the pronunciation of an aspirated sound of the initial letter⁸. However, in various Punjabi dialects, we find a variety of pronunciations, such as in the Majhi of Central Punjab, the words written with Pairin Haha would certainly be pronounced with a high tone, however, in most other dialects (both Western Punjabi and Eastern Punjabi dialects), either a complete or seminal /h/ would be found, or in places we would find the aspirate sound⁹.

In the Old Gurmukhi of the Sri Guru Granth Sahib, both Udaat and Pairīn Haha have been used. This is a result of the wide range of Punjabi dialects, apart from other languages, being represented in Gurbani, at different stages in their evolution (from the 12th century to the 17th century). The Udaat suggests the high tone, while the Pairīn Haha denotes the aspirate /h/ with the inherent vowel being suppressed. In modern Gurmukhi, only Pairīn Haha is used, but orthographically it does not suggest the high tone. Both high tone and /h/ pronunciation are to be found among the Punjabi dialects.

The Halant or Virama of Devanagari, which has the similar form to Udaat, is used in English-Punjabi dictionaries to transcribe the correct pronunciation of English words and in other technical writings, such as lexicons.

It is recommended that Udaat be encoded as a separate Unicode character, with the following properties:

```
0A51;GURMUKHI SIGN UDAAT;Mn;0;NSM;;;;;N;;;;;
```

The UTC may wish to change the canonical combining class as required.

Udaat differs very slightly in its graphical appearance when compared to Halant. Udaat starts with a small tip and slopes inward to the right whereas Halant has a more uniform thickness and slopes outwards to the right.



Udaat should push down U and UU in the same way that existing subjoined consonants do.



⁵ The *Punjabi-English dictionary*, published by Punjabi University, Patiala (1994) gives following meanings of the term Udaat: 'sublime; acutely accentuated, sharply intoned' (p. 9)

⁶ Kulbir S Thind, *Text Trivia* in Gurbani-CD 2004. The basis of the file is the Sri Guru Granth Sahib, published by the Shiromani Gurdwara Parbandak Committee in 1994.

⁷ Harkirat Singh, *Gurbani di Bhasha te Vyakaran* (1997, in Punjabi), pp. 102-3.

⁸ Joginder Singh Talwara, *Gurbani da Saral Viakarn-Bodh*, part I, pp. 27-8.

⁹ Ibid, p. 103.

Udaat should be placed after the consonant whose tone is being changed but before the vowel. In many ways, **Udaat should be treated as a subjoined consonant**. In the following examples, an acute accent indicates the high tone.

ਖੋਲ੍ਹਿਓ (Khōlī 'ō)

0A16	0A4B	0A32	0A51	0A3F	0A13
ਖ	ੋ	ਲ	੍	ਿ	ਓ

ਸੰਮਾਲੇਹਾਂ (Samhālēhām)

0A38	0A70	0A2E	0A51	0A3E	0A32	0A47	0A39	0A3E	0A02
ਸ	ੰ	ਮ	੍	ਾ	ਲ	ੇ	ਹ	ਾ	ੰ

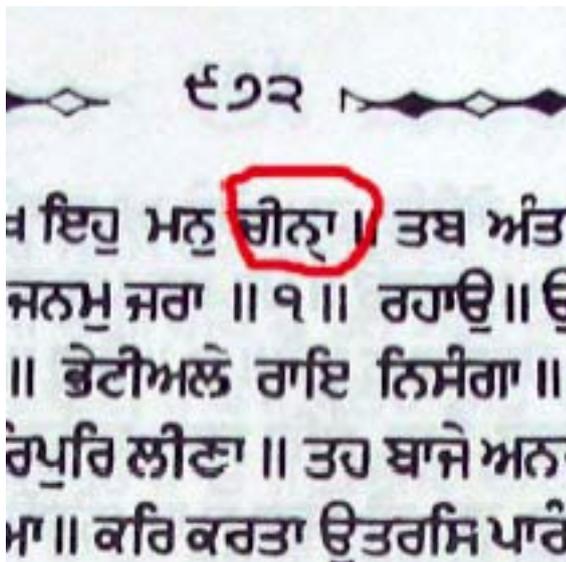
ਓਲਾਮੇ (Ōlāmhē)

0A13	0A32	0A3E	0A2E	0A51	0A47
ਓ	ਲ	ਾ	ਮ	੍	ੇ

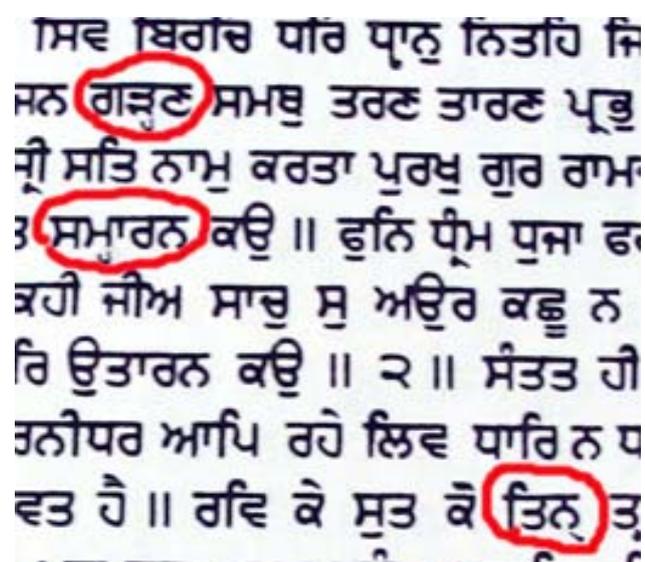
Usage

The use of Udaat is limited to a few texts, namely the Sikh holy book and other religious literature. This should not negate its need for encoding – especially for Sikhs who wish to encode their holy book in Unicode without altering the original form of the text.

Evidence for its use is provided:



D1 SGGS page 972 (੯੭੨)



D2 SGGS page 1404 (੧੪੦੪)

Evidence of the use of Udaat may not be prevalent in other texts. However the officially sanctioned copy of the Guru Granth Sahib uses the character a total of 1188 times.

E1. Proposal Summary

A. Administrative

1. Title

Proposed Changes to Gurmukhi 3

2. Requester's name

Sukhjinder Sidhu (Punjabi Computing Resource Centre)

3. Requester type (Member body/Liaison/Individual contribution)

Individual contribution.

4. Submission date

2005-10-27

5. Requester's reference (if applicable)

6. Choose one of the following:

6a. This is a complete proposal

Yes.

6b. More information will be provided later

No.

B. Technical – General

1. Choose one of the following:

1a. This proposal is for a new script (set of characters)

No

1b. The proposal is for addition of character(s) to an existing block

Yes.

1c. Name of the existing block

Gurmukhi

2. Number of characters in proposal

1

3. Proposed category (see section II, Character Categories)

Category C

4a. Proposed Level of Implementation (1, 2 or 3) (see clause 14, ISO/IEC 10646-1: 2000)

Level 1

4b. Is a rationale provided for the choice?

No

4c. If YES, reference

5a. Is a repertoire including character names provided?

Yes.

GURMUKHI SIGN UDAAT

5b. If YES, are the names in accordance with the character naming guidelines in Annex L of ISO/IEC 10646-1: 2000?

Yes.

5c. Are the character shapes attached in a legible form suitable for review?

Yes.

6a. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?

Dr K Thind, True Type

6b. If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:

Development version of AnmolUniBani available by request by emailing sukhuk@users.sourceforge.net

7a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?

Yes

7b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?

Yes

8. 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)?

No.

9. 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.

Yes. See above.

C. Technical – Justification

1. Has this proposal for addition of character(s) been submitted before? If YES, explain.

Yes, an incomplete proposal was submitted in "Proposed Changes to Gurmukhi" (L2/05-088).

2a. 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.

2b. If YES, with whom?

Jeevan Deol

Manudeep Singh
Serjinder Singh
Kulbir Thind
And others

2c. If YES, available relevant documents

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?

No.

4a. The context of use for the proposed characters (type of use; common or rare)

Common (Archaic)

4b. Reference

5a. Are the proposed characters in current use by the user community?

No.

5b. If YES, where?

6a. After giving due considerations to the principles in Principles and Procedures document (a WG 2 standing document) must the proposed characters be entirely in the BMP?

Yes..

6b. If YES, is a rationale provided?

Yes.

6c. If YES, reference

Additional Gurmukhi characters.

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

No.

8a. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?

No.

8b. If YES, is a rationale for its inclusion provided?

8c. If YES, reference

9a. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?

Yes.

9b. If YES, is a rationale for its inclusion provided?

Yes.

9c. If YES, reference

Yes, see A1. Compatibility with existing conventions.

10a. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?

Yes.

10b. If YES, is a rationale for its inclusion provided?

Yes

10c. If YES, reference

See C1

11a. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC10646-1: 2000)?

Yes.

11b. If YES, is a rationale for such use provided?

Yes.

11c. If YES, reference

See above.

12a. Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?

No.

12b. If YES, reference

13a. Does the proposal contain characters with any special properties such as control function or similar semantics?

No.

13b. If YES, describe in detail (include attachment if necessary)

14a. Does the proposal contain any Ideographic compatibility character(s)?

No.

14b. If YES, is the equivalent corresponding unified ideographic character(s) identified?

E2. Appendix

ਸੁਣਹਿ ਗਾਵਹਿ ਪਰਭਾਤਾਹ
ਸਹਿਤ ਕਰੰ ॥ ਕੰਚਨੁ ਤਨੁ ਹੋਇ
ਲ ਮਹਿ ਰਹਿਆ ਪੂਰਿ ਬਹੁ ਬਿ
ਰਣੈ ॥ ੪ ॥ ਜਿਨਹੁ ਬਾਤ ਨਿਸੁਲ
ਕ ਮਹਿ ਜਲਹਰ ਬਿੰਬ ਜੁਗਤਿ
ਸਿਰੀ ਗੁਰੂ ਸਾਹਿਬੁ ਸਭ ਉਪਰਿ
ਤੇ ॥ ਕਵਲ ਨੈਨ ਮਧੁਰ ਬੈਨ ਕੋਟਿ
ਮਤਿ ਅਨਪ ਮੋਹ ਮਹਾ ਮਗ ਭਈ

E1 SGGS page 1402 (੧੪੦੨)

ਨਯਣ ॥ ੧ ॥ ਸਤਿ ਰੂਪੁ ਸਤਿ
ਸਤਕਿ ਧੁਰਿ ॥ ਪ੍ਰਗਟ ਜੋਤਿ
ਰੂ ਕਹਾਯਉ ॥ ਭਨਿ ਮਥੁਰਾ ਮੁ
ਗੁਰੂ ਸਗਲ ਸਿਸ੍ਰਿ ਲਗਿ ਬਿਤਰ
ਕੈ ਹਿਤੁ ਨਾਮ ਸਿਉ ॥ ਪ
ਸਿਉ ॥ ਅਪਰ ਪਰੰਪਰ ਪੁਰਖ
ਕੋ ਪ੍ਰਭੁ ਸੂਬ ਮਯ ਅਰਜਨ ਰ

E2 SGGS page 1408 (੧੪੦੮)

ਉ ॥ ਰਤਨ ਨਾਮੁ
ਹਰਿ ਗੁਰੁ ਨਾਨਕ
ਵਨ ਬਿਧਿ ਉਸਤਿ
ਸੇਖ ਗੁਣ ਅੰਤੁ ਨ
ਲ ਪ੍ਰਭਿ ਸੇਇ ਜਨ
ਭ ਦਾਤਉ ਦਾਤਾਰ
ਦਾਤਿ ਪਦਗਾਇ

E3 SGGS page 1386 (੧੩੮੬)

ਸਦਾ ਗੁਰ ਕਾ ਘਾਟ ਘਾਟ ਪਰ
ਹਿ ਇਸ਼ਾਨੁ ॥ ਇਸ਼ਾਨੁ ਕਰਹਿ
ਸਿ ਪਾਰਸ ਕਉ ਜੋਤਿ ਸਰੂਪੀ
ਬਰਨੰ ॥ ਸਤਿਗੁਰੁ ਗੁਰੁ ਸੇਵਿ ਅ
ਮ ਸਾਨੀ ਤੇਈ ਜੀਵ ਕਾਲ ਤੇ ਬ
ਗੁ ਰਚਾ ॥ ਕੁੰਡਲਨੀ ਸੁਰਝੀ
ਬਚ ਕ੍ਰਮ ਸੇਵੀਐ ਸਚਾ ॥ ੫ ॥
ਨ ਧੰਗ ਧੰਗ ਕਹਾਰ ਮਾ ਕਸੋਦ ਨਿ

E4 SGGS page 1402 (੧੪੦੨)

ਸਰਿ ਸੰਤੋਖ ਸਮਾਇਯਉ ॥
ਨਾ ਬਦਨਿ ਬਰ ਦਾਤਿ ਅਲਖ
ਹਜਿ ਨਿਜ ਘਰਿ ਸਹਾਰਉ ॥
ਨ ਕਲੁਚਰੈ ਤੈ ਜਨਕਹ ਕਲਾ
ਲੈ ਨਗੇ ॥ ਨੇਜਾ ਨਾਮ ਨ
ਗੀ ਕਾ ਬੋਹਿਥਾ ॥ ਤੁਅ ਸਤਿ
ਗਰ ਕਨੈ ਪਾਇਅਉ ॥ ਅਸ

E5 SGGS page 1408 (੧੪੦੮)