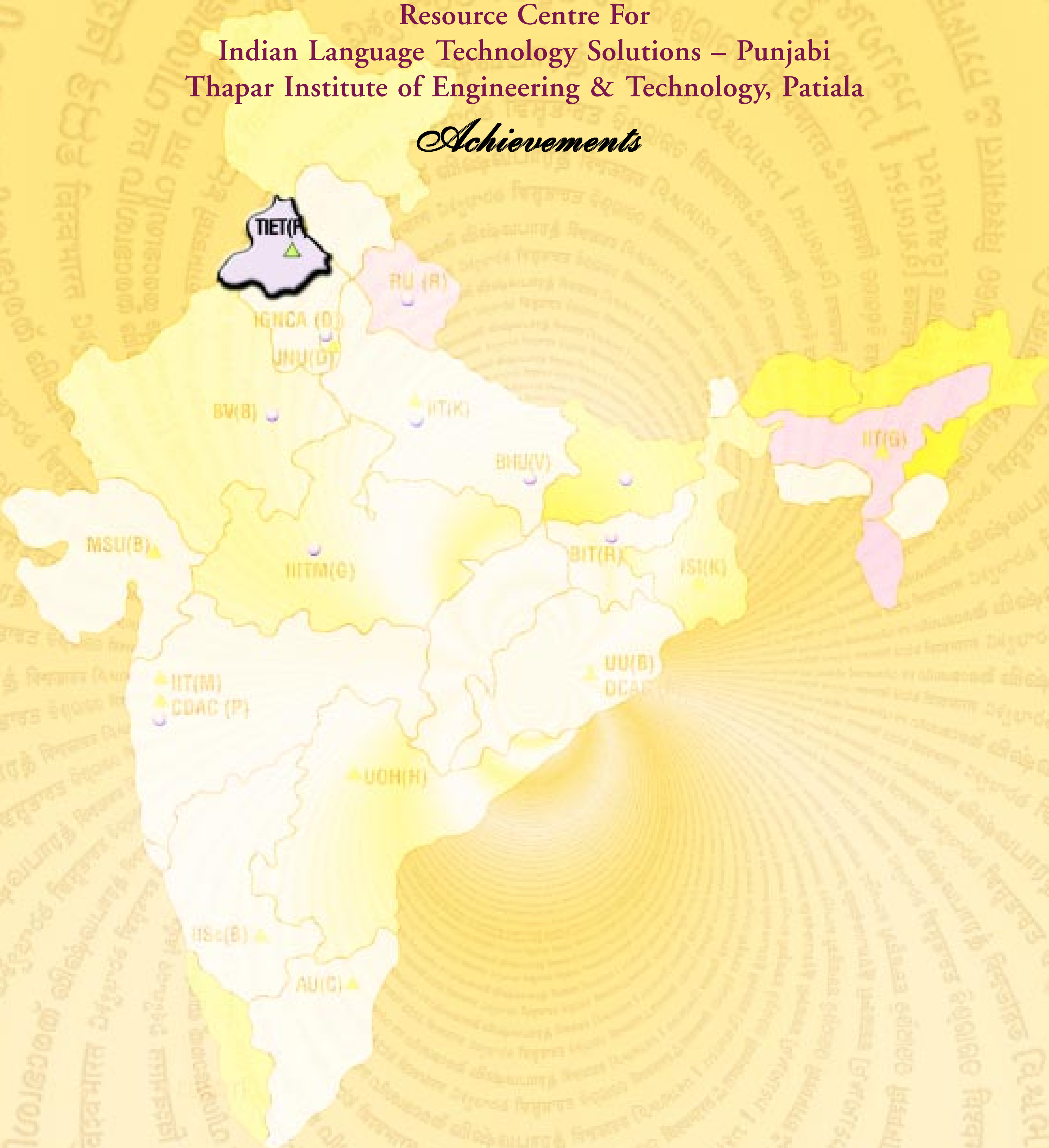
Resource Centre For
Indian Language Technology Solutions – Punjabi
Thapar Institute of Engineering & Technology, Patiala

Achievements



School of Mathematics and Computer Applications (SMCA)
Thapar Institute of Engineering & Technology
(Deemed University) Patiala-147001
Tel. : 00-91-175-2393382 E-mail : punjabirc@yahoo.com
Website : http://punjabirc.tiet.ac.in

RCILTS-Punjabi
Thapar Institute of Engineering & Technology, Patiala

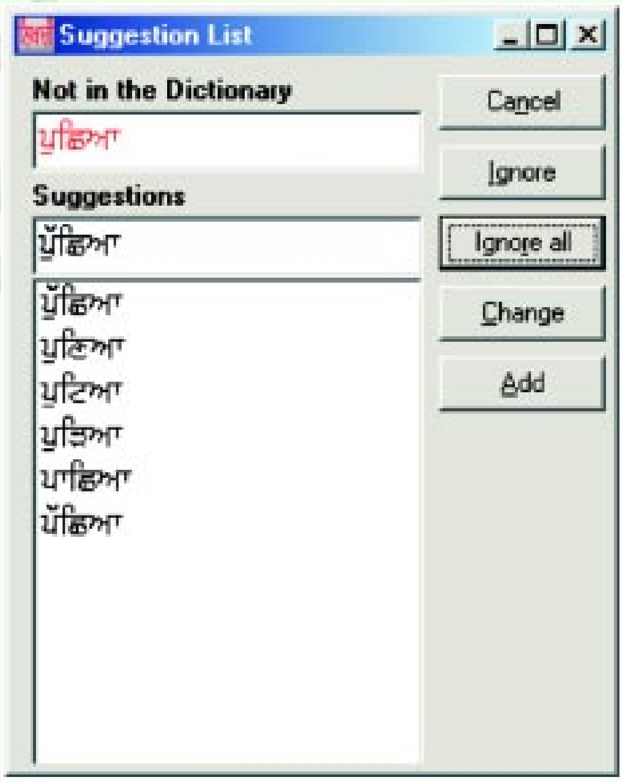
Introduction

The Resource Centre for Indian Languages Technology Solution-Punjabi was established in April 2000 at Thapar Institute of Engineering & Technology working in TDIL Programme of Department of Information Technology. The Resource Centre is supported by the Ministry of Communications and Information Technology (MCIT) with the aim of working for technology development of Punjabi and also to provide access to millions of people across the globe. The main task before the resource centre has been the promotion of Punjabi language so as to extend the benefit of knowledge and awareness across the globe. It has since evolved into a multi-faced research center. Before the Punjabi Resource Centre was set up, very little work had been done for computerization of Punjabi, even though large number of Punjabis settled in USA, UK and Canada have been using computers for long. The only work done was development of Punjabi fonts and some Punjabi websites. There was no Punjabi spell checker, Punjabi sorting utility, electronic dictionaries or Gurmukhi OCR. In the span of three years, we have developed for Punjabi lexical resources, content creation tools, Gurmukhi OCR and uploaded Punjabi text on web. The details of the work done during the last three years is as follows:

1. Products Developed

1.1 **Spell Checker** : A spell checker is a basic necessity for composing text in any language. Till now, no spell checker was available for Punjabi as a result people had to waste lot of time proof reading the Punjabi text. This problem was more severe for publishers and writers residing abroad who used computers to type Punjabi material, as even Punjabi dictionaries are not easily available to them. To make the matters worse, there has been no standardization of Punjabi spellings. In fact there are numerous words which are spelled in more than one way, some are being written in 3-4 different ways. After discussions with Punjabi scholars it was decided to retain multiple spellings for the commonly used words and a spell checker for Punjabi has been developed taking Harkirat Singh's *Shabad Jor Kosh* as the base and additional words were taken from the dictionaries published by Punjabi University and the State Language Department and the corpus developed by CIIL, Mysore.

For generating the suggestion list, a study was conducted to discover the most common errors made by Punjabi typists. A list of similar sounding words and consonants was also complied and a suggestion list using this knowledge was generated based on reverse edit distance. It was found that the right suggestion is presented as a default suggestion in majority of the cases. In such a case, the user needs only to confirm the default suggestion and proceed with the next error. Otherwise, the user needs to scroll through a list of



ਪੁਛਿਆ

ਪੁੱਛਿਆ

ਪੁੱਛਿਆ
ਪੁਚਿਆ
ਪੁਟਿਆ
ਪੁੜਿਆ
ਪਾਛਿਆ
ਪੱਛਿਆ

Fig.1 : Spell Cheker

suggestions and pick one as the right one. The spell checker supports text typed in any of the popular Punjabi fonts as well as ISCII encoded files. The Punjabi spell checker is now complete and MOU is soon going to be signed with M/S Modular Systems, for Transfer of Technology of the spell checker. Output of the Spell Checker is shown in Fig. 1.

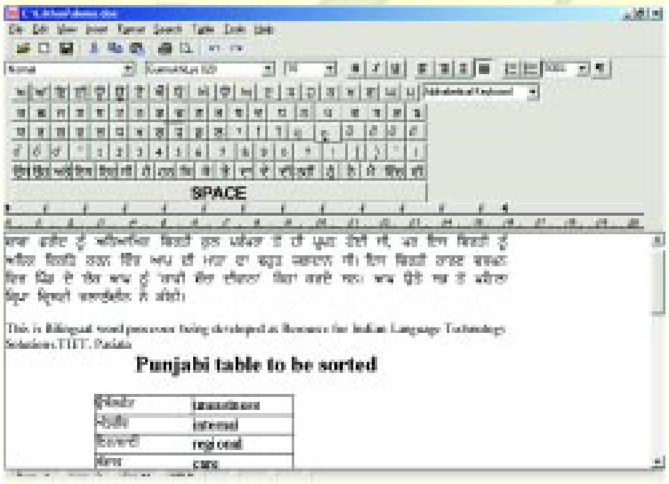
1.2 **Font Converter :** The major development in computerization of Punjabi before the setting up of the Punjabi Resource Centre was creation of Punjabi fonts. As a lot of Punjabis reside in USA, Canada and UK, so a large variety of fonts were developed. In the absence of any standardization, each font had its own keyboard layout. The net result is that now there is a chaos as far as the Punjabi language in electronic form is concerned. Neither can one exchange the notes in Punjabi as conveniently as in English language, nor can one perform search on texts in Punjabi. This is so because the texts are being stored in font dependent glyph codes and the glyph coding schemes for these fonts is typically different for different fonts. To alleviate this problem, a font conversion utility has been developed at the Resource Centre. The utility supports more than sixty different Punjabi fonts and thirty two keyboard layouts. A user may type text in any font and using this utility he can later on convert it to any other font. This utility will be a great help to publishers, writers and people exchanging text in different fonts.

1.3 **Sorting Utility :** Sorting and Indexing is one of the basic necessities of the database management system such as maintenance of student's records or arranging a dictionary in alphabetic order. But unfortunately there does not exist any software for automatic sorting of Gurmukhi words and all such work has to done manually. The collating sequence provided by UNICODE or ISCII is not adequate, as it is not compatible with the traditional sorting for Gurmukhi words. Gurmukhi, like other Indian languages has a unique sorting mechanism. Unlike English, consonants and vowels have different priorities in sorting. Words are sorted by taking the consonant's order as the first consideration and then the associated vowel's order as the second consideration. In addition there is another complication. The properly sorting "characters" in Gurmukhi often requires treating multiple (two or three) code points as a single sorting element. Thus we cannot depend on character encoding order to get correct sorting instead we have to develop using sorting rules of Gurmukhi, collation function which converts the word into some intermediate form for sorting. After discussions with linguists and studying in detail the alphabetic order in Punjabi we have developed a general sorting algorithm which works on text encoded in any popular Gurmukhi font or ISCII.

1.4 **Bilingual Punjabi/English Word Processor**: A Bilingual Punjabi/English Word processor has been developed keeping in view the difficulties being faced by the users while working on the currently available word processors. It was found that as far as Punjabi is concerned no word processor which addresses the word processing requirements typical for Punjabi language, such as Punjabi spell checker, Punjabi thesaurus and dictionary, virtual Punjabi keyboards and Punjabi sorting and font conversion utilities, has been developed. In this direction, efforts were made a Punjabi word processor **Likhari** was developed. **Likhari** supports word processing under the windows environment and allows typing and processing in Punjabi Language through the common typewriter keyboard layout. It has MS-Word compatible features and commands. It provides a number of features that make the use of Punjabi Language on a computer easy and provides a number of tools to increase the efficiency of the user. These tools include

Bilingual Spell Checker with suggestion list, Onscreen keyboard layouts with composition reference for Punjabi Language typing, bilingual search and replace, sorting as per the language, alphabetical order, technical glossaries and onscreen Bilingual dictionaries. The main features of **Likhari** are :

- Very simple user interface
- Online active Keyboard for users who do not know how to type in Punjabi.
- Choice of Phonetic, Remington and Alphabetic Keyboard layouts with composition reference.
- Bilingual Spell Checker for Punjabi and English.
- Bilingual Search and Replace.
- Support for sorting the text in English or Punjabi as per the language alphabetical order.
- Support for more than 60 commonly used Punjabi fonts.
- Support for features like Tables, Numbering, Bullets, Character & Paragraph formatting, Page set up, Print Preview, Header and Footer etc.



[illegible]

This is Bilingual word processor being developed at Resource for Indian Language Technology Solutions, TIET, Patiala

Punjabi table to be sorted

[illegible]	[illegible]
[illegible]	internal
[illegible]	regional
[illegible]	care

Fig 2 : A Screen Shot of Punjabi WordProcessor Likhari

- Online Technical English-Punjabi glossary.
- Support for. ISCII, .TXT, DOC, .RTF and .HTML file formats.
- Extensive help at various levels to make it easy for the user to learn.

1.5 Gurmukhi OCR : Optical Character Recognition (OCR) is the process whereby typed or printed pages can be scanned into computer systems, and their contents recognized and converted into machine-readable code. The text, which the machine can read, has great advantages over that merely displayed as an image, since it can be edited, exported to other programs and indexed for retrieval on any word. If one needs to electronically store and manipulate large amounts of text or printed matter such as newspapers, contracts, letters, faxes, price lists or corpus development, one will find OCR programs can save lots of effort. For the first time a complete OCR package for Gurmukhi script has been developed. The OCR has a recognition accuracy of more than 97%. The OCR can automatically detect and correct skewed pages and can also detect page orientation. The OCR can recognize multiple Punjabi fonts and sizes. The main features of the Gurmukhi OCR are:

Recognition Accuracy

Recognition accuracy is around 97% for books, photocopies and medium degraded documents and around 99.5% for laser print outs and good quality documents.

Import Image

B/W Images, 24-bit color images, 256 grayscale images.

Output File Format

ISCII / text file encoded in any one of the popular Punjabi fonts.

Fonts Supported

All non-decorative Gurmukhi fonts.

Proofing

On-screen verifier

Additional Features

Inbuilt Spell checking facility. Automatic skew detection and correction (skew range – 5 to +5 degrees). Upside down image auto detection and correction.

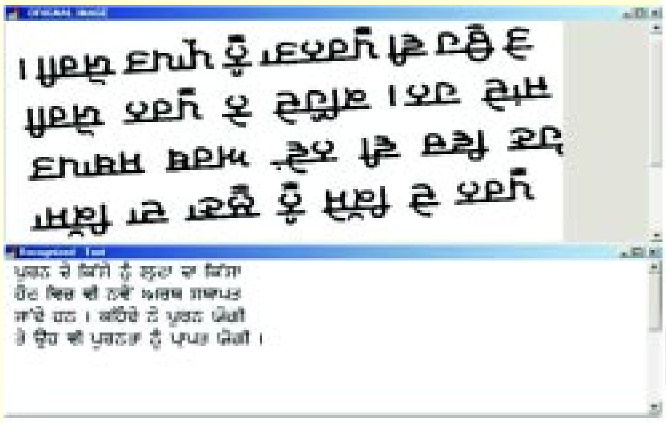


Fig. 3 : A Screen shot of Gurmukhi OCR

1.6 Gurmukhi to Shahmukhi Transliteration:

Punjabi language is used in both parts of Punjab in India and Pakistan. In East Punjab (India) Punjabi is written in Gurmukhi script. This script was invented by Guru Angad Dev ji. This is written from left to right. In West Punjab (Pakistan) Punjabi is written in Persian script , also known as Shahmukhi script. This is written from right to left like Urdu and Persian. In East Punjab many Sanskrit words are absorbed into Punjabi language and similarly in West Punjab many Persian and Urdu words are used in Punjabi. But the essence of Punjabi language remains same. Both Shahmukhi and Gurmukhi have been in use simultaneously. After the Partition, Punjabi on the Indian side of the border was restricted officially to the *Gurmukhi* script. On the Pakistani side it was restricted to *Shahmukhi* script. The result is that a script-wall has come up between the two sides of the Punjab, which prevents cultural and literary exchanges — there is ignorance on both sides about developments in contemporary prose and poetry on the other side. To break this wall, it is necessary to develop transliteration programs which can automatically convert Gurmukhi text to Shahmukhi and vice-versa. In this direction the Punjabi Resource Centre has collaborated with the Urdu Resource Centre established at CDAC, Pune to develop a computer program which automatically converts Gurmukhi text into Shahmukhi. Using this software, a collection of short stories penned by Mr. K. S. Duggal in Gurmukhi has been converted into Shahmukhi (Figs 4-5).

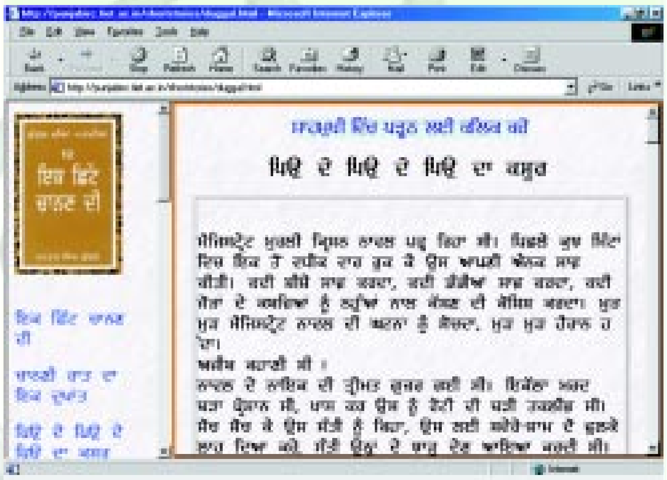


Fig 4 : A Short Story in Gurmukhi

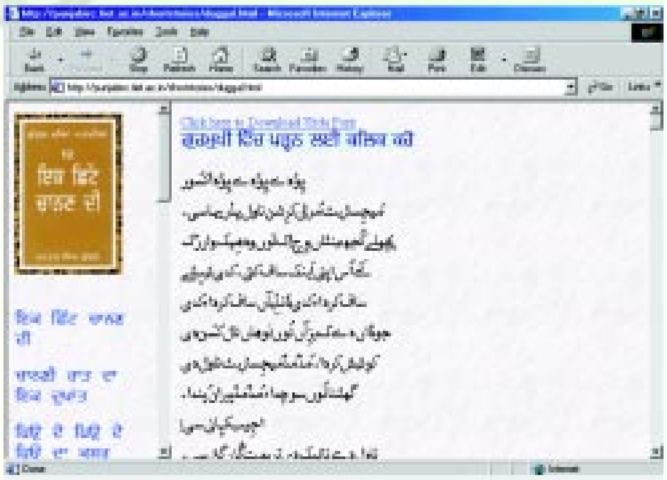


Fig 5 : The Short Story in Gurmukhi of Fig. 4 converted to Urdu

2. Contents Uploaded on Internet

For benefit of common man, Punjabi literature, dictionaries and products for free download have been uploaded on the Resource Centre's website (http://punjabirc.tiet.ac.in). The contents uploaded on the internet are:

2.1 **Literature :** Punjabi Classics such as *Bullehshah Dian Kafian, Farid De Salok, Heer. Luna, Chandi di War, Japji Sahib* have been uploaded in Punjabi using Punjabi Dynamic Fonts. The website contains the detailed description of these classics. Tool tips of difficult words have also been provided in Punjabi. Audio clips of these classics have also been included.

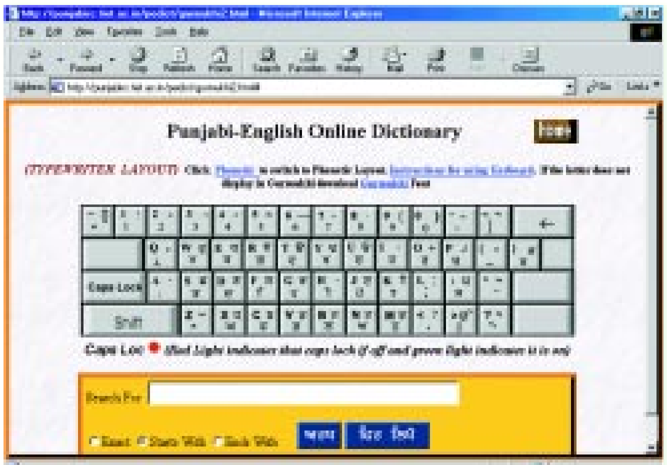
2.2 **Bilingual Dictionaries and glossary :** The following bilingual dictionaries and glossary have been uploaded at our site.

- **Punjabi English On-line Dictionary :** A Punjabi English dictionary is made available on the website. The dictionary has about 40,000 Punjabi words. The dictionary has sample sentences for common words and audio clips of pronunciation of Punjabi words have also been provided. The Dictionary is accessible in both typewriter layout and phonetic layout. One can search for complete match or a pattern in the dictionary.

- **English Punjabi On-line Dictionary:** An English-Punjabi dictionary containing about 45000 words has been made available on the website.

- **Hindi Punjabi On-line Dictionary:** A Hindi-Punjabi dictionary containing about 45000 words has also been made available on the website.

- **Glossary of English-Punjabi administrative terms :** A technical glossary of around 17,000 English-Punjabi administrative terms has also been



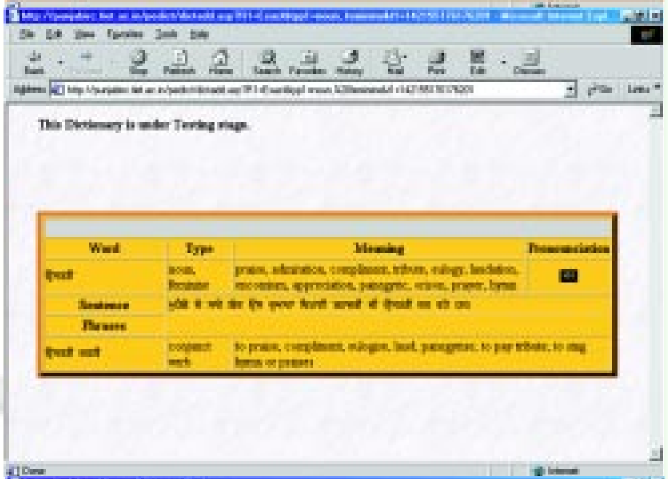
Punjabi-English Online Dictionary

(TYPEWRITER LAYOUT) Click Phonetic to switch to Phonetic Layout. Instructions for using Keyboard. If the letter does not display in Gurmukhi download Gurmukhi Font

Caps Lock (Red Light indicates that caps lock if off and green light indicates it is on)

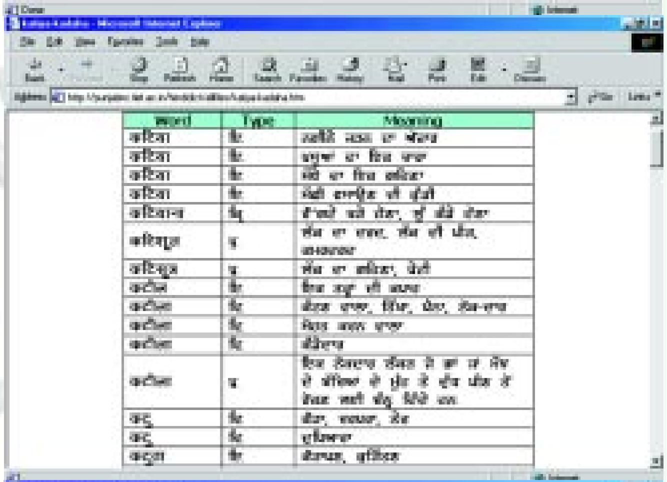
Search For:

○ Exact ○ Starts With ○ Ends With

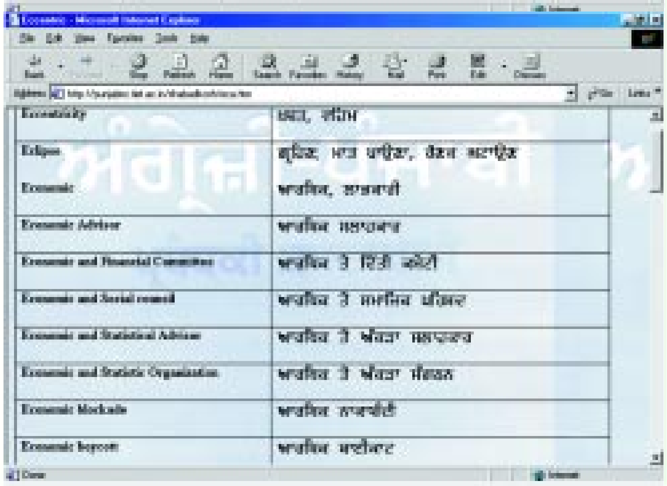


This Dictionary is under Testing stage.

Word	Type	Meaning	Pronunciation
[illegible]	noun, feminine	praise, admiration, compliment, tribute, eulogy, laudation, encomium, appreciation, panegyric, orison, prayer, hymn	
Sentence	[illegible]		
Phrases			
[illegible]	conjunct verb	to praise, compliment, eulogize, laud, panegyrize, to pay tribute, to sing hymns or prayers	



Word	Type	Meaning
[illegible]	[illegible]	[illegible]



Eccentricity	[illegible]
Eclipse	[illegible]
Economic	[illegible]
Economic Adviser	[illegible]
Economic and Financial Committee	[illegible]
Economic and Social council	[illegible]
Economic and Statistical Adviser	[illegible]
Economic and Statistic Organisation	[illegible]
Economic blockade	[illegible]
Economic boycott	[illegible]

uploaded. A CD has also been developed and the glossary has been installed on computers in many offices of Punjab State Government.

2.3 **On Line Teaching of Punjabi :** For benefit of Punjabis settled abroad and others interesting in learning Punjabi, a web site for On-line teaching of Punjabi has also been developed. Work is complete for Gurmukhi Orthography, Punjabi Pronouncing rules and a limited vocabulary both in text and pictorial format along with audio effects is also provided. The alphabets can be learnt in Animation pattern as we draw with hands.

2.4 **On Line Font Conversion Utility :** An online font conversion utility has been provided. The user can paste Punjabi text encoded in any of the supported sixty Punjabi fonts and convert it to the desired Punjabi font.

2.5 Downloads

- **Punajbi Spell Checker –** A Punjabi-English spell checker has been uploaded on the home page and any user can download and install the spell checker on his system.

- **Punjabi Fonts –** Two Punjabi fonts(LIKHARI_P and LIKHARI_R) for phonetic and keyboard layouts have been developed and made available for free download. The fonts are available in both True Type and Dynamic formats.

3. Interaction with Punjab State Government

A five day training programme was organised for staff of State Language Department. They were given training in word processing, email and internet. We have been in constant touch with the Secretary IT, Punjab and a demonstration of the products developed at our RC was given to him. The Punjabi wordprocessor and Gurmukhi OCR were installed on the systems in the office of Secretary, IT, Punjab. We have been providing technical support to State Language Department. Besides providing training to their staff, we have also set up their web site in Punjabi. A CD of English-Punjabi administrative terms has been developed for the Sate Language Department and the CD has been installed in Patiala DC office and State Revenue office.

4. Publications

1. G.S.Lehal, "State of Computerization of Punjabi", *Proceedings Second World Punjabi Conference,* Prince George, Canada (2003). (Accepted for Publication)

2. G.S.Lehal, "A Gurmukhi Collation Algorithm", *Journal of CSI*. (Accepted for publication)

3. G.S.Lehal, Chandan Singh and Renu Dhir, "Structural feature-based approach for script identification of Gurmukhi and Roman characters and words", *Document Recognition and Retrieval X*, Proceedings SPIE, USA, Vol. 5010 (2003). (Accepted for publication)

4. G.S.Lehal and Chandan Singh, "A complete OCR system for Gurmukhi script", *Structural, Syntactic and Statistical Pattern Recognition*, T. Caelli, A. Amin, R.P.W. Duin, M. Kamel and D.de Ridder (Eds.), Lecture Notes in Computer Science, Vol. 2396, Springer-Verlag, Germany, pp. 344-352, (2002).

5. G.S.Lehal, Chandan Singh and Renu Dhir, "Automatic separation of Gurmukhi and Roman script words", *Proceedings Indo-European Conference on Multilingual Communication Technologies,* Pune, R. K. Arora, M. Kulkarni and H. Darbari (Editors), Tata McGraw-Hill, pp. 32-38,(2002).

6. G.S.Lehal and Chandan Singh, "A Post Processor for Gurmukhi OCR", SADHANA *Academy Proceedings in Engineering Sciences,* Vol. 27, Part 1, pp. 99-112, (2002)

7. G.S.Lehal and Chandan Singh, "Text segmentation of machine printed Gurmukhi script", *Document Recognition and Retrieval VIII*, Paul B. Kantor, Daniel P. Lopresti, Jiangying Zhou, Editors, Proceedings SPIE, USA, Vol. 4307, pp. 223-231, (2001).

8. G.S.Lehal and Chandan Singh, "A technique for segmentation of Gurmukhi text", *Computer Analysis of Images and Patterns,* W. Skarbek (Ed.), Lecture Notes in Computer Science, Vol. 2124, Springer-Verlag, Germany, pp. 191-200, (2001).

9. G.S.Lehal, Chandan Singh and Ritu Lehal, "A shape based post processor for Gurmukhi OCR", *Proceedings of 6th International Conference on Document Analysis and Recognition*, Seattle, USA, IEEE Computer Society Press, USA, pp. 1105-1109, (2001).

10. G.S.Lehal and Nivedan Bhatt, "A recognition system for Devnagri and English handwritten numerals", *Advances in Multimodal Interfaces – ICMI 2001,* T. Tan, Y. Shi and W. Gao (Editors), Lecture Notes in Computer Science, Vol. 1948, Springer-Verlag, Germany, pp. 442-449. (2000).

11. G.S.Lehal and Chandan Singh, "A Gurmukhi script recognition system", *Proceedings 15th International Conference on Pattern Recognition*, Barcelona, Spain, IEEE Computer Society Press, California, USA, Vol 2, pp. 557-560, (2000).

5. The Team Members

Dr. R. K. Sharma	rksharma@mail.tiet.ac.in
Dr. G. S. Lehal	gslehal@mail.tiet.ac.in
Dr. Rajesh Kumar	rakumar@mail.tiet.ac.in
Rajeev Kumar	rkumar@mail.tiet.ac.in
Ramneet Mavi	ramneet_mavi@rediffmail.com
Deepshikha Goyal	deepshikha26us@yahoo.com
Nivedan Bhatt	nivedanbhatt@rediffmail.com
Sukhwant Kaur	sukhwant2010@yahoo.co.in
Ramanpreet Singh	shallukalra@yahoo.co.in
Parneet Cheema	NA
Pallavi Dixit	NA
Rupsi Arora	NA
Yoginder Sharma	NA
Pooja Dhamija	NA
Zameerpal kaur	NA
Dr. Kuljeet Kapoor	NA
Dr. Devinder Singh	NA
Karamjeet Kaur	NA
Jaspal Singh	NA
Manpreet Singh	NA
Rakesh K. Dawra	NA
Shallu Kalra	NA
Kuldeep Kumar	NA
Baljit Singh	NA
Aarti Gupta	NA
Sunita Sharma	NA
Surjit Singh	NA

Courtesy: Prof. R.K. Sharma
Thapar Institute of Engineering & Technology
Department of Computer Science & Engineering
Patiala 147 001
(RCILTS for Gurmukhi)
Tel: 00-91-175-2393137, 393374, 2283502
E-mail : rksharma@mail.tiet.ac.in

RCILTS – Punjabi