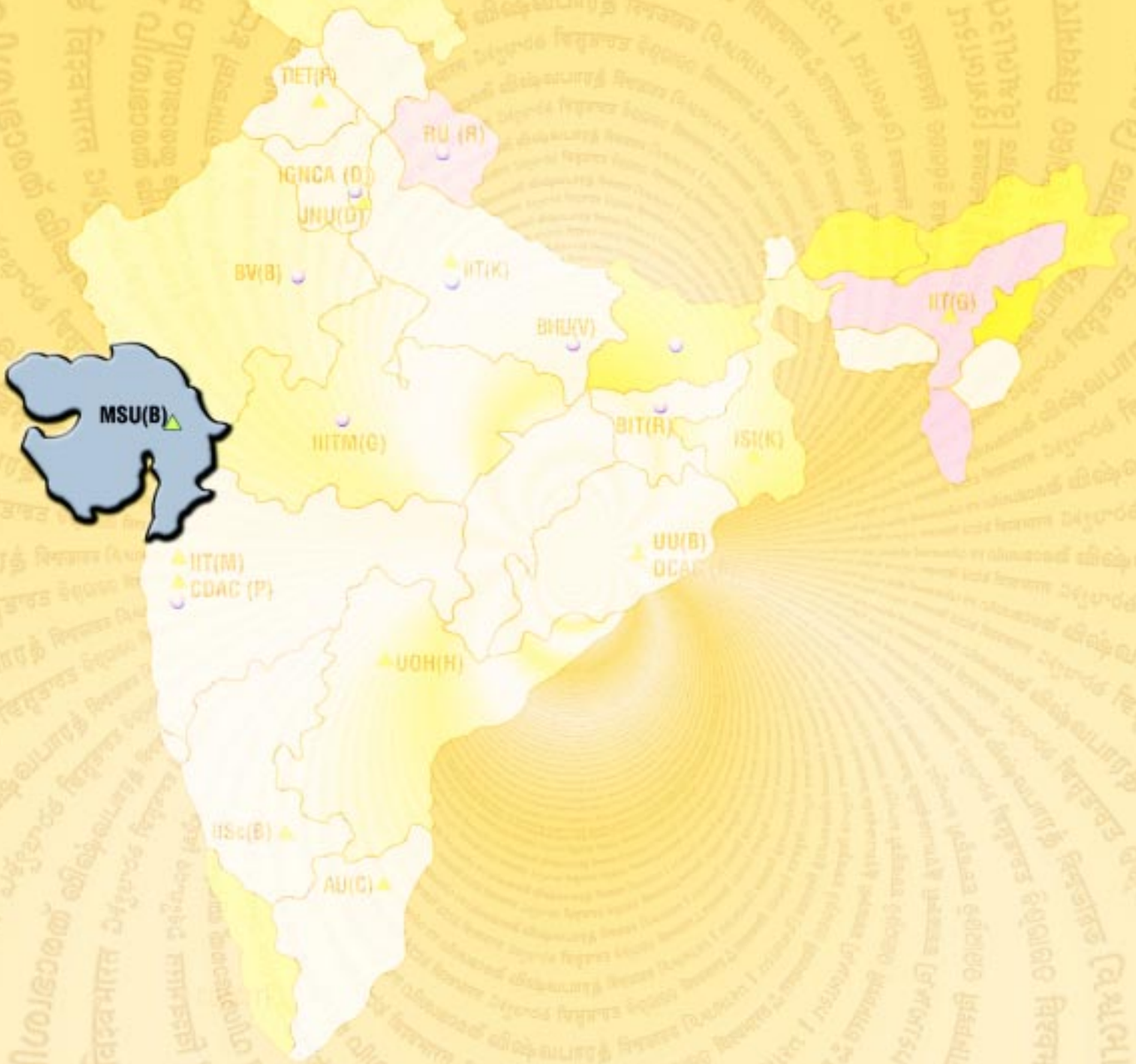
Resource Centre For
Indian Language Technology Solutions – Gujarati
The Maharaja Sayajirao University of Baroda, Vadodara

Achievements



Department of Gujarati, Faculty of Arts
The Maharaja Sayajirao University of Baroda, Vadodara-390002, India
Tel. : 00-91-265-2792959 E-mail : rciltg@satyam.net.in
Website : http://msubaroda.ac.in/rciltg/

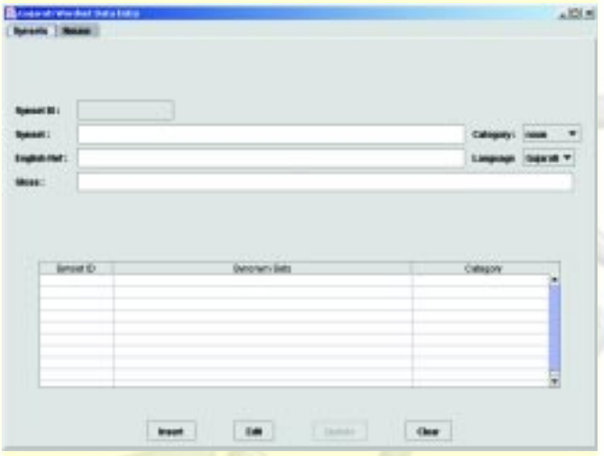
RCILTS-Gujarati
Maharaja Sayajirao University, Baroda

1. Knowledge Resources

(Corpora, Parallel Corpora, Multi Lingual Digital Dictionaries)

1.1 Gujarati WordNet (Lexical Resource)

Gujarati WordNet, a Lexical Resource, is in the process of development. RC at MSU is working in close collaboration with and through able guidance of Professor Pushpak Bhattacharya of RC at IIT, Mumbai. The Data Entry Interface for entering the synsets is build. Presently above 500 synsets are already been developed.



Further work on preparing the synsets and entering into the database is going on. Basic Semantic Relationships are also been established like Hypernymy, Hyponymy between the synsets formed and entered in the database. Further work is going on for establishing other Semantic relationships between the synsets. The Web Interface is already under construction to facilitate the access of an On-Line Lexical Database of Gujarati to be available over the Internet.

2. Knowledge Tools

(Portals, Language Processing Tools, fonts, morphs analyzer, spell checker, text editor, basic word processor, code conversion utility, etc.)

2.1 Portal

RC for Gujarati has a web site in English hosted at http://msubaroda.ac.in/rciltg/ . The Gujarati version of the same will be available soon. The web site contains the information about the software developments at RCILTG. The full text of the poems and prose of well known Gujarati authors is planned to be displayed on the web site. A manuscript of 19th Century classic "Saraswatichandra" by Govarthanram Madhavram Tripathi, in the author's own handwriting is located and digitized and would be put on the web for the interested researcher soon.

2.2 Multilingual Text Editor

Unicode based Multilingual Text Editor which was required since long for Indian Languages, has been developed at the RC for Gujarati.

The important feature of this Text Editor is that it fulfils the necessity of Universal storage format (Unicode) for Indian Languages. The Indian languages that are given support by this Text Editor are Bangla, Gujarati, Gurmukhi, Hindi, Kannada, Malayalam, Marathi, Oriya, Sanskrit, Tamil and Telugu along with English as International language.

To make typing ease, assistance is given to the user by providing an onscreen INSCRIPT keyboard layout for all languages supported. The Text Editor supports the basic file features like creating new file, opening existing file and saving file. It is efficient enough to open file of around 64KB. Also the basic text editing features are provided like cut, copy, paste, undo, redo, find, find next, replace and select all.

The user can use any of the system fonts for writing text. The styles that are supported by Text Editor are font name, font size, bold, italic and underline. The sorting facility is also available based of word sort or line sort. The support for three different Look and Feels viz., Motif, Windows and Metal of Java are provided to the user. The Gujarati Spell Checker is also bundled with this Text Editor.

Current Focus

• To increase the efficiency of Text Editor.

Further Plans

• E-mail support for Indian Languages.

• ISCII compatible Import/Export facilities.

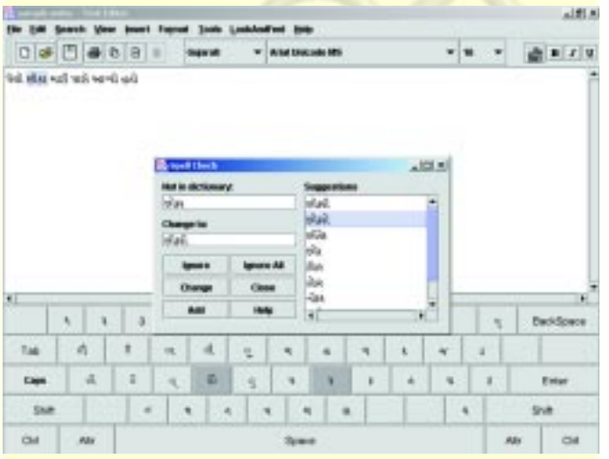
• More formatting features to be provided.

2.3 Code Converter

We have the programs which can convert the codes between Unicode and ISCII.

2.4 Gujarati Spell Checker

The Gujarati Spell Checker is in its last stage of development. It is based on a Morphological Analysisof Gujarati language in order to increase the intelligence of the Gujarati Spell Checker. Morphological Analyzer covers the analysis of Nouns and Verbs of the language.



The suggestion generation is very effective. The number of suggestions that will be generated will not be greater then ten. All the non-gujarati words along with different symbols are ignored and are not checked. The Gujarati Spell Checker is integrated along with Multi-Lingual Text Editor for the convenience to user.

Currently the software is under testing and very soon the accuracy in percentage of Morphological Analysis will be available.

Current Focus

- Rigorous testing to be done, using large corpus, to find all the faults in the spell checker and morphological analyzer.
- Improving the algorithms for the correctness and efficiency of the spell checker.

Further Plans

- Increasing the size of the root dictionary covering the maximum possible words of the language.
- Making an independent spell checker which can be used on Unicode or ISCII compatible text.
- More and more testing to increase the correctness of the spell checker and also improving the suggestion generation.

3. Translation Support Systems

(Machine Translation, Multi Lingual Information Access, Cross Language Information Retrieval)

3.1 Machine Translation

The work on this front has already started and initially we have planned to be domain specific

and we have selected weather related text. A group of two is working on this area. Basic Universal Word Lexicon format is being studied.

Out of all the approaches we have selected to go for Interlingua (UNL) based approach for this purpose. The team has started the exploring the various resources available on the UNDL Foundation website. The specifications for enconverter and deconverter are being studied. The UNL proxy has been downloaded and tested with the sample documents.

4. Human Machine Interface Systems

(Optical Character Recognition Systems, Voice Recognition Systems, Text to Speech Systems)

4.1 OCR for Gujarati

A working prototype of Gujarati OCR which is capable of doing end to end task i.e. converting an image into an equivalent text, has been produced. The OCR employs the template matching technique, the same as the Telugu OCR developed at RC for Telugu uses, for recognition and nearest neighbour for classification. The Gujarati OCR differs in many ways from the Telugu one due to the special features like the necessity for zone separation.

The program, which can be invoked from command line, takes a gray image, scanned in color or grayscale, in any of the popular image formats like JPEG, GIF, PNG, BMP and after processing it produces the output as Plain Text file in Unicode format. The input image is assumed to be skew-corrected, without any images and tables, having only one column. We have a converter which coverts Unicode to equivalent ISCII and vise-versa ready, so we can produce the ISCII output as and when required.

Rigorous testing of the OCR for recognition rate etc. is being initiated and the initial results are quite good. Work is in progress for fine tuning various modules in order to improve the recognition accuracy.

Current Focus

- In the test we have seen some misrecognition, so we are currently testing the performance of Discrete Cosine Transform as a feature extractor.
- The programs are getting reorganized to in the Object Oriented way.

Future Plans

- Developing/Adapting skew-correction algorithm.
- Developing the post-processing routines.
- Thorough testing of the OCR and necessary fine tuning to take it to the required accuracy of 97% or above.
- Integrating with the OCRized spell-checker.
- Developing GUI for the system with inbuilt Text Editor.

4.2 Text To Speech (TTS)

The work on TTS has already been started at RC-Gujarati. A team of three students has explored the available TTSs such as FreeTTS, IBMJS in English and *Dhvan*i and *Naarad* in Hindi. All of these are the implementation of Java Speech API specifications. We have made a significant progress also in this direction.

While starting this work, we had divided our task in the following two major subtasks :

1. Developing a program which generates the output, which can be input to the existing speech engine. E.g. IBMJS takes phonetic transcription of the text using IPA codes, to produce the sound.
2. Developing a core speech engine with native accent. We are adopting the concatenation based approach.

Task Completed So Far

- A program is now available with RC-Gujarati which allows user to type in Gujarati text, using INSCRIPT keyboard layout and then it converts the string into its equivalent IPA string and feeds it to the IBM speech engine, which in turn produces the sound represented by given IPA string.

Current Focus

- Exploring the phonological properties of language, extracting the basic phonemes from a continuous speech and design of basic speech engine.

Future Plans

- Development of preprocessor to handle language specific details.
- Developing speech engine and integrating with preprocessor.

5. Localization

(Adapting IT tools and solutions in Indian language(s)- IT localisation clinic interaction with state government for possible solutions in Ils).

5.1 Language Technology Human Resource Development

(Manpower Development in Natural Language Processing - specialized training programs)

RC for Gujarati is trying to develop man power by providing the resources to final year students from the relevant background like Mathematics, Computer Science, Electrical Engineering, Electronics Engineering, Computer Application etc. for doing their projects with RC. Two (one ME (Electrical) and one M.Sc. (Statistics)) students have already completed their projects on study and development of various techniques for Gujarati OCR.

At presents Five students are working for their final year projects. Three of which are working on TTS technologies and two are working on development of UNL based machine translation support system.

Apart from this RC, Gujarati has trained three physically challenged persons in Gujarati data entry using inscript keyboard and now they are doing the data entry for RC and for which they are being paid.

6. Standardization

(Contribution towards UNICODE, Transliteration Tables(s), XML, lexware format, etc)

6.1 UNICODE

UNICODE for Gujarati have been studied and some additions and corrections have been suggested. They are as follows :

- Correction in the shape of AVAGRAH
- Addition of Gujarati Currency sign, Symbols for ¼, ½ , ¾ .
- Addition of two rendering rules for combination of Ja with Dependent vowel sign E and EE

The collation sequence for Gujarati alphabets is not same as the sequence of the characters in the UNICODE chart. Hence the correct collation sequence have been finalized.

All the above modifications and suggestions were published in the April,2002 issue of Vishwabharat magazine of TDIL.

In addition to above the RC for Gujarati has worked closely with Mr. Abhijit Dutta of IBM, Delhi and helped him in standardizing on the collation sequence for Gujarati and the set of Glyphs for developing the fonts.

Apart from development of basic technology RC for Gujarati has, as per the original proposal accepted by the Ministry, developed a strong Knowledge base in Gujarati, available on multimedia CDs and on our website. This knowledge base includes the following;

1. Select Gujarati classics in print and manuscripts, including a rare mss of the 19th century classic, Sarasvatichandra in the author's handwriting, with many changes he has made incorporated in the digitized form on CD.

2. Cultural History and Geography of Gujarat on CD, useful to researchers in one set and to tourist in a second set.

3. Multimedia CDs on Rural Health Problems and Preventive Care.

4. Language self-learning CDs useful to students (learning Sanskrit, English etc.) and second generation NRIs and new Researchers abroad.

7. Products which can be launched and the service which the Resource Centre can provide to the State Government and the Industry in the region.

7.1 Technical Skills

The RC for Gujarati has built technical skills related to Image Processing and Pattern Recognition, Audio signal processing, Speech Synthesis, Unicode based text processing and rendering and implementation of these using Java. These can readily be shared, with proper agreement, with researches and industry in the region. This knowledge can even be transferred to the new generation by teaching some courses in these areas at the nearby colleges.

RC-Gujarati has developed Multilingual Text Editor which can be used to type in Unicode. With the help of this, we could propagate the use of standard encoding storage of data in Unicode or ISCII. In turn this would lead to the situation where the data can be transferred from one machine to the other without transferring the fonts. This could be much use in Government offices where such data transfer is a daily routine.

7.2 Products Which can be Launched

At present RC has a multilingual text editor ready to launch. The detail of which is given above.

7.3 IT Services

RC for Gujarati has developed the following services which it can provide to the State Government and the Industry in the region:

- Multimedia CDs on Language Self-learning ("Learn Gujarati through English", "Learn Sanskrit through Gujarati", "Learn English through Gujarati" etc.) could be of use to learners in Gujarat and overseas, including the NRIs. In opur interaction with him, these CDs have been found of interest to Dr. M.N. Cooper of Infosis Ltd. Pune.

- Multimedia CDs on Rural Health and Adolescent Heath, developed in collaboration with Tribhuvandas Foundation, Karamsad and its Director, Dr. Nikhil Kharod, M.D., could be of use to the Departments of Health and Education, Gujarat Government and to the Panchayats.

- Services on Library Knowledge-base, giving detailed information on books in print, 19th and 20th century journals and medieval manuscripts

available at some libraries in Gujarat and Mumbai, on CDs and our Website, could be of much use to Universities, Colleges and other institutions of research and teaching in India and abroad. The

libraries with which we have entered into service agreements have undertaken to provide information through email, fax and surface mail to users who would have access to this knowledge base through our Website.

8. IT Services - Multimedia CDs

8.1 Learn Gujarati Through English

A multimedia tutorial is designed that helps an enthusiast, learning Gujarati language though English. The tutorial is based upon the concept of Dr. Jagadish Dave who teaches Gujarati in Europe. He has used this methodology to teach Gujarati to the persons who doesn't have any connection to Gujarat and for that matter India.

The main features of the CD are:

- Novel approach for teaching Gujarati by exploiting similarity of shapes for remembering various characters.
- Starts with a familiar graphic resembling a Gujarati letter and shows the manufacture of related letters from it.
- For example, starting with the symbol 'S' which is a character in English as well as Gujarati, the Gujarati characters with related shapes *DA, KA, PHA, HA, THA* are derived as morphological modifications of the basic symbol.
- The related audio commentary figuratively explains the modifications of the basic symbol.
- The pronunciation of Gujarati words in which the character is embedded at various positions also accompanies the presentation.
- After introducing a set of alphabets the learner is lead to meaningful word formation involving those alphabets.
- Introduction of grammar through simple sentences instead of complicated rules.

Future Plans

- A practical conversation involving those words suitable for handling a real life situation is to be provided at a regular intervals
- At the end of the presentation a list of common words, their meanings and their pronunciation is to be added.
- The less common grammar rules are to be added.
- A video of some practical situations is to be incorporated.

8.2 Adolescent Health

RCILTG organized a two days workshop on Adolescent Health with Tribhuvandas foundation of Anand on the above topic. About 40 teachers participated in the workshop. The workshop was aimed at adolescent health and how teachers can scientifically disseminate the knowledge about the changes occurring in male and female body during this age.

This CD is a gist of ideas and suggestions, came out after a few long sessions of brain-storming. The information presented here is the work of Dr. Nikhil Kharod (M.D. Pediatrician).

The features of this CD are:

- Introduction of sensitive issues from a scientific view point.
- Accompanying pictures, sketches, diagrams and tables giving medically proved information about the topic.
- A detailed description of nutrients, their source and disease caused in case of their deficiency.
- Description of the Body cycle of male and female.
- A quiz with key answers which gives, the user, and an idea about his /her knowledge in this area.

Our future plan includes the production of a series of learning language CDs. Work on the CD "Learn Sanskrit through Gujarati" has already begun. Next to follow would be a set of CDs linking Gujarati with other Indian languages and with English. There is a good marketing possibility for this series.

Initial work has been completed on production of a set of CDs with multimedia presentation on Cultural History and Cultural Geography of Gujarat. This series would be useful for Primary and Secondary Schools and also for tourism.

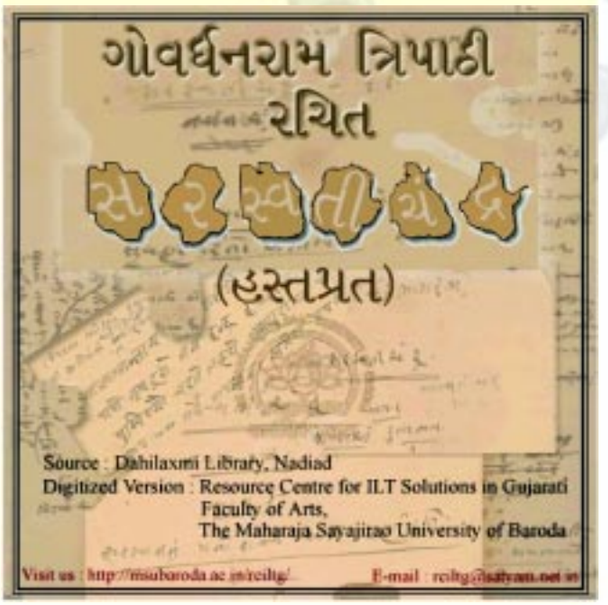
One CD on "Rural Health" has been produced, focusing on Adolescent Health. Similar workshops on other areas on Rural Health would be conducted to produce more CDs on other aspects of health in rural Gujarat.

8.3 Bibliography of Books

In close collaboration with some important research libraries of Gujarati and Mumbai, the RC for Gujarati has worked on preparing on-line bibliography of books and journals available at these centres. Also on-line lists of critical and creative items (essays, fiction, poetry, drama, culture-related information) published in 19th and early 20th century Gujarati journals, has been digitized. The Libraries have agreed to provide photo-copied of the knowledge-base items, at no-profit level, to the users of the RC's website.

8.4 Sarasvatichandra

Original manuscript of the 19th century classic, **Sarasvatichandra**, by Govardhanram Tripathi, which had a great impact on cultural life of Gujarat, has been located by our RC. The mss, in the author's handwriting, has now been digitized and made into a CD with an appropriate navigation.



Note: RC for Gujarati has trained three physically challenged former students, Kum. Rina Phadia, Shri Ganpat Solanki and Arun Paladgaonkar, in data entry and has used their paid services for its data entry work for Wordnet and CD production.

9. The Team Members

Prof. Sitanshu Mehta	pi_rcmsu@satyam.net.in
Prof. S. Rama Mohan	srm-cc@msubaroda.ac.in
Mr. Jignesh Dholakia	jignesh-rcilt@msubaroda.ac.in
Mr. Mihir Trivedi	mihirktrivedi@rediffmail.com
Mr. Irshad Shaikh	ishaikh16@yahoo.com

Courtesy: Shri Sitanshu Y. Mehta
M.S. University of Baroda
Department of Gujarati
Faculty of Arts
Baroda – 390 002
(RCILTS for Gujarati)
Tel : 00-91-265-2792959
E-mail : rciltg@satyam.net.in