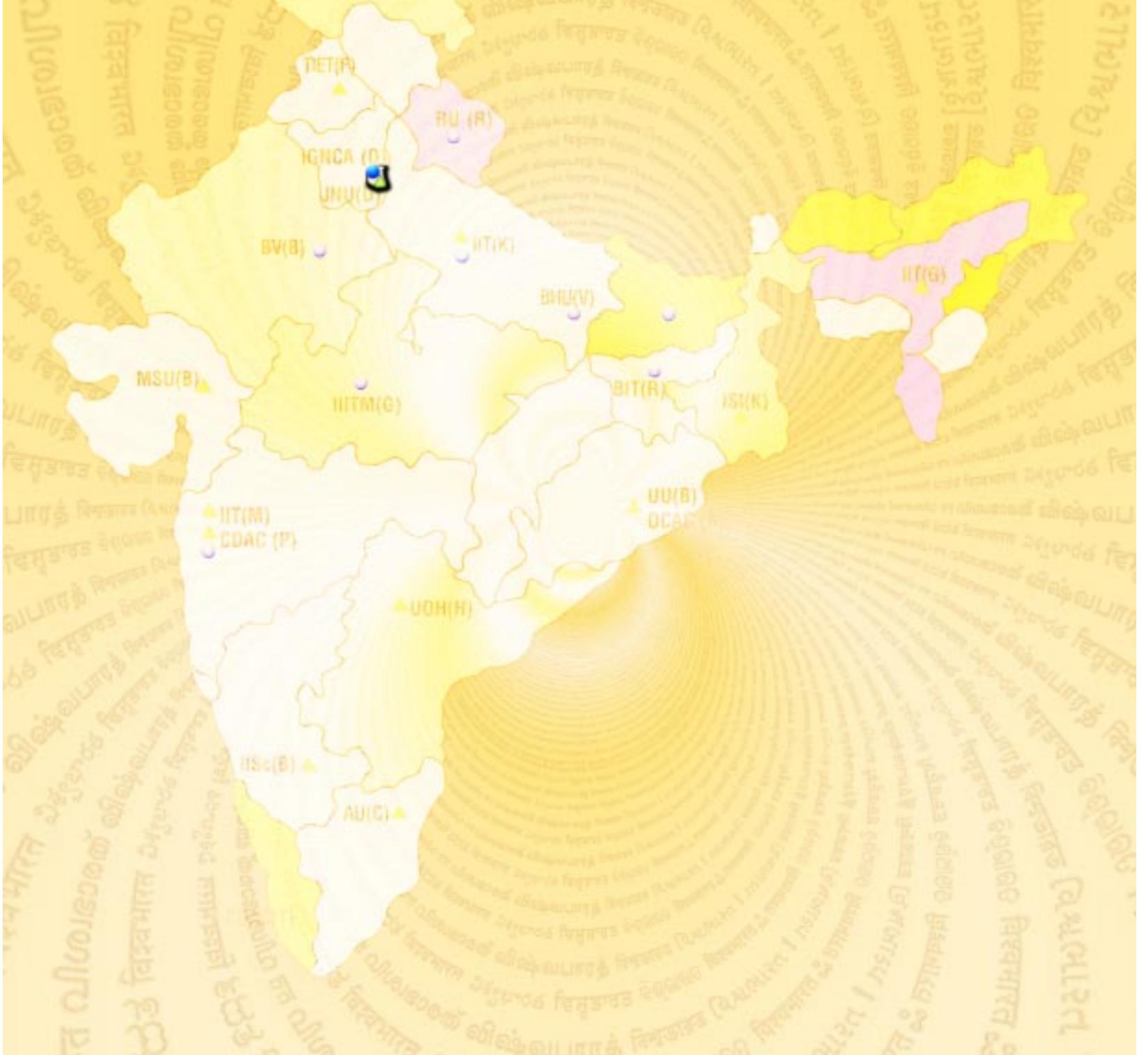


Resource Centre For
Indian Language Technology Solutions – Sanskrit, Japanese, Chinese
Jawaharlal Nehru University, New Delhi

Achievements



Jawaharlal Nehru University, New Delhi-110067

Tel. : 00-91-11-26704772

E-mail : gvs10@hotmail.com

Website : <http://202.41.10.205/rciltsjnu>



**RCILTS-Sanskrit, Japanese, Chinese
Jawaharlal Nehru University, New Delhi**

Introduction

RCILTS, JNU is a resource center for Sanskrit language of DIT, Government of India. At JNU work started in three languages viz., Sanskrit, Japanese, and Chinese. It took lot of time and effort to assemble a cohesive team for the development of language technology.

In the beginning lot of problems were faced regarding development tools, software professionals, as well as language professionals. Due to the socio-economic nature of the IT industry the attrition of the IT professionals was very high, so work suffered a lot. Till today it is becoming very difficult to get good linguists (computational linguistics) who could carry forward the task of future development.

In the first stage it was decided that we should not be reinventing the wheel. We shall try to develop language technology and resources that is not being addressed at other RCILTS. India has rich cultural heritage and time proven scientific knowledge. All this is largely in the form of Sanskrit literature. So it was decided that we would develop a web based Sanskrit Language Learning System (as a second language primarily). It would be of great use to those scholars who look to our heritage knowledge for designing and developing *Knowledge* based systems.

Keeping the above goal in mind at RCILTS, JNU, we have designed various modules that teach Sanskrit language in an *Asynchronous* fashion. In the process of development of these language resources we have kept in mind the various standardization aspect of creating these resources. Here we would like to mention specially the Unicode. All the web content that is available on RCILTS JNU site is Unicode compliant.

In the early phases of development we used lot of tools developed at C-DAC. So lot of content was in ISFOC/ISCII based. Earlier all the processing was done on ISCII data. There was no tool where we could create content that were Unicode compliant. At a very later stage of development when Unicode tools/techniques were developed and few of them bought, then we could convert all the content in

Unicode. During this conversion process we developed *Web Content Unicode Converter*. This converter works with ISFOC/ISCII data.



The RCILTS, JNU website is available at <http://202.41.10.205/rciltsjnu>

The various modules that have been developed at RCILTS JNU, is presently integrated for learning the language. With the development of new modules this language learning system shall be used for various kinds of language processing (text and speech). The various software modules and language resources that are available are

Learning materials

- *Sanskrit Lessons*
- *Sanskrit Exercises*
- *Sanskrit Script Learning System*

Lexicons

- *Sanskrit – English Lexicon*
- *Sanskrit – English Lexicon of Nyaya terms*
- *English – Sanskrit Lexicon*

Sanskrit Grammar

- *Panini's Asthadhyayi*
- *Laghu Sidhyant Kaumudi*
- *Dhatu Roop to Dhatu*
- *Dhatu to Dhatu Roop*
- *Pratipadik Roop*
- *Shabd Roop*
- *Upsarg*
- *Sandhi*

Tools

- *HTML Content Converter*

- ISCI to UNICODE Database Converter and vice versa
- Devnagari Unicode Keyboard Driver for Browsers

Sanskrit Grammar is primarily based on *Panini's Ashtyadhyayi* and *SidhyantKaumudi*. We have already developed modules for *Dhatu Roop*, *Sabd Roop*, and *Sandhi*. After learning all these modules a learner knows the basics of Sanskrit language.

We plan to develop *Sandhi Vichhed* System. With the development of *Sandhi Vichhed* system as tool it shall be very easy for the scholars to read and understand the age old Sanskrit manuscripts which are treasures of time proven scientific knowledge. With the proper understanding of these manuscripts it would be easy for them to interpret the scientific knowledge in their original context and apply them in the present context for the development of knowledge based software systems. Knowledge based Systems would aid in the process of Administration, Business, Management, and Education.

Since JNU has a rich language school, in the beginning work was also carried out in the area of Japanese and Chinese. But after some timework on Chinese language discontinued due to paucity of Chinese language professionals. But work in the Japanese language is still on. We have developed lot of language learning resources for Japanese too.

1. Sanskrit Lessons and Exercises

Sanskrit Lessons are useful to learn Sanskrit Language. There are 21 lessons and the number will increase. Lessons are designed so that, a person who doesn't know Sanskrit at all but knows either English or Hindi can go through these lessons and learn. Sanskrit contents of the Lessons are described both in English and Hindi.

The basics of the Sanskrit Language are covered elaborately. And then the grammar part of the Sanskrit language is explained. Nearly all topics of grammar are covered with lot of examples and explanation. Special care is taken for the pronunciation of the Sanskrit alphabets and the words. The pronunciation is based on standard phonetic notations. One can go to the 'Sanskrit script learning module' from the Lesson to learn the writing process and pronunciation of alphabets.



Some extra facilities are given to learn and remember easily. All the Sanskrit words, used through out the lessons are available to the user with their roots and meanings. These words are also connected to the Lexicon and accessible from any page of any lesson. User can also see the new words on a particular page that he is reading with the facilities explained above. The Navigation Buttons bellow the each page is self-explanatory.

The exercises are based on the lessons. For each Lesson more than two exercises are available. These are objective type questions. Four options are given for each question. As the option is chosen, result is given promptly to the user that it is correct or not. The Correct answers of a particular question will be given in the table bellow after attempting the questions. So user can see the answer that he has given as well as the actual answer of that question just choosing an alternative. Exercises are obviously helpful to practice what a user learns on that Lesson.



2. Devanagari Script learning and Character Recognition System

Scripts are the basics of a Language. And writing of these scripts are very important to learn a Language. Using this module the user learns Devnagari alphabets and learns how to write and pronounce them. Writing strokes for a character are shown with stroke orders with four different speeds. It takes care of all characters, conjuncts, matras and numerals also. And there is audio support to listen the pronunciation of the alphabet. Recognize a character after hearing its pronunciation is also over. It is also hyper-linked with Sanskrit Lessons. User can access this module from Sanskrit Lesson.



Example- The Sanskrit Vowel “Riri” is written in four stroke with four different speed. Except Basic alphabets there are numbers of conjuncts characters in Sanskrit Language. Their formation and the writing process in the proper stroke orders are given. Phonetic notations are given in the left-top corner so that a user can pronounce the alphabet easily.

Numerals are also important to learn as their frequent use in the language. Writing process of Sanskrit numeric numbers are shown with their stroke order. If there is more than one style of writing of numeric numbers then the alternatives are provided bellow.

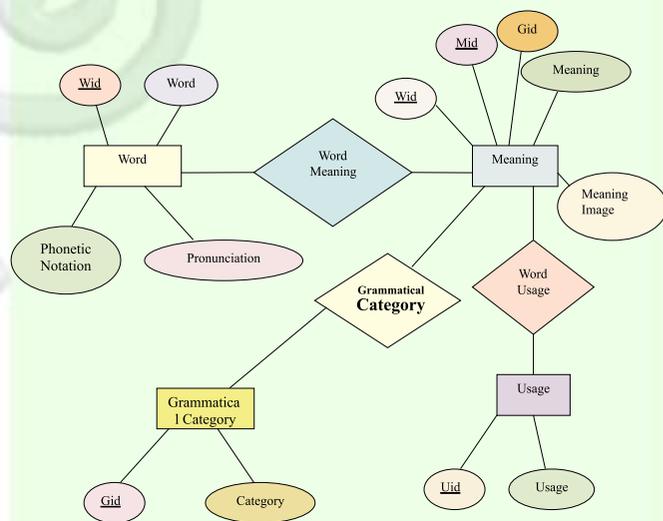
Pronunciation of alphabet is important to speak a language. The attention is given for the right and clear pronunciation of the alphabets and audio supports are provided. There are some exercises also for the users in recognizing alphabets according to the pronunciation.



3. Sanskrit-English Lexicon

In this module a user can get the meaning, its grammatical category and usages of *Sanskrit* word. It will also display the phonetic notation of the word. Grammatical category and usage can be displayed if available.

The E-R Diagram of Sanskrit English Lexicon is given below. All the Lexicons follow the same E-R Diagram



The facility for getting input from the user is provided in both ways. Users can click the beginning letter of the word. It will display all words, which start with that letter in the right side window. On clicking the particular word the meaning with phonetic notation and the grammatical category are displayed in the left window.



User can write the input word with the help of Phonetic Keyboard in the text box. On pressing the submit button the meaning with phonetic notation and the grammatical category are displayed for that inputted word.

4. Dictionary of Nyaya Term

This dictionary of Nyaya terms is a part of an ambitious project of preparing an encyclopedia of Indian logic. The main purpose is to resolve the confusion and make uniformity in translating the Nyaya terms in English. The stress has been given to the clarity of concepts than on literal translation. The root and meanings with phonetic notation of Sanskrit Nyaya term is provided here.

For giving input, it is necessary to click on the beginning letter of the word in the left window. It will display all words, which are start with that letter in the right side window. On clicking the particular word the root and meanings with phonetic notation are displayed in the left window.



5. Panini Astadhyayi

In this module a user can get the *Panini Sutra*, *Sutra with Anuvrtti*, its *Sanskrit Explanation* and *English Explanation* after giving *Astadhyayi Sutra* number or *Siddhanta Kaumudi Sutra* number. The user will also get the information that how many time the *Astadhyayi Sutra* occur in the *Siddhanta Kaumudi*.

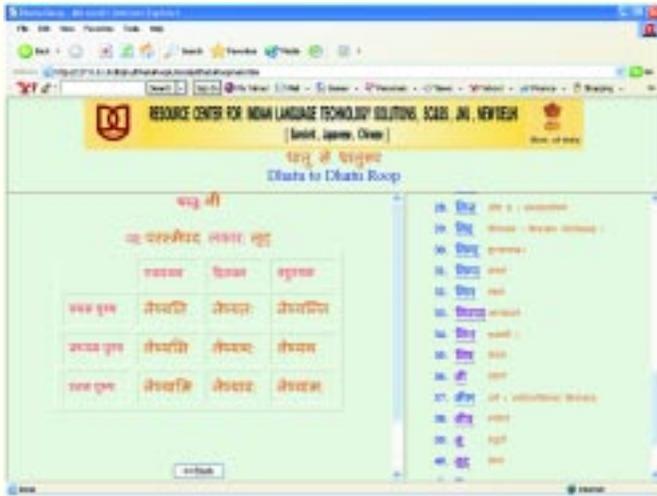
The user can select *Astadhyayi Sutra* or can enter the *Siddhanta Kaumudi* Number (If the user gives both *Astadhyayi Sutra* and *Siddhanta Kaumudi* Number then the system will accept *Astadhyayi Sutra*).

Example- For the inputted *Astadhyayi Sutra* Number 2.2.23, this *Sutra* occurs in two places (*Kaumudi* Serial Number 529 and 829). If the user clicks any of the *Kaumudi Sutra* Number then all the information of that *Sutra* will be displayed. If the user clicks *Kaumudi Prakaran* then all the sutra of that *Prakaran* will be displayed.



6. Dhatu to DhatuRoop

Using this module a user can get all the information of a *dhatu*, i.e. a user can get *dhatu's* meaning, its *gan* name, *pad* name and the *roops* after selecting the *lakar*. For giving input, it is necessary to click on the beginning letter of a word in the left window. It will display all words, which are started with that letter in the right side window.



Example- The inputted *dhatu* 'ni' and the information of the *dhatu* 'ni' is displayed. The 10 *lakar* name is displayed the user can select any *lakar* and get the *roops* of that *lakar*. As this *dhatu* is an *Ubhyapada Dhatu*, the user has to select the *pad* also in the next page.

7. Dhatu Roop to Dhatu

This module is the reverse of the *Dhatu to DhatuRoop* module. In this module a user can get the *dhatu* and its attributes from a *dhatu roop*.

Example- For the given DhatuRoop "Bhavati", User can get all the informations related to that Dhaturoop like Dhatu Name, Dhatu Arth, Dhatu Gana, Dhatu Pada, Dhatu Lakar, Purush and Vachan.

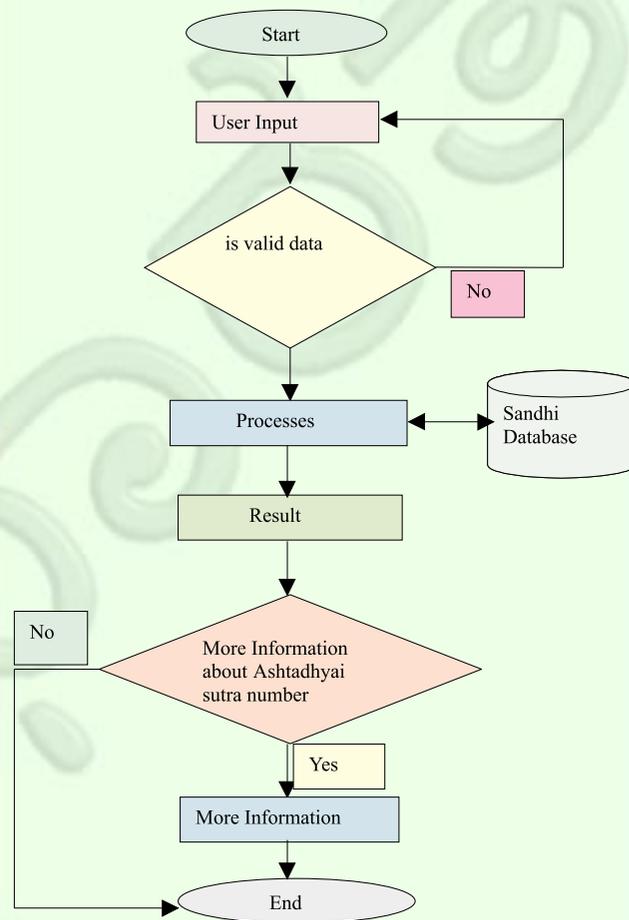


8. Sandhi

Sandhi means the coalescence of two words coming in immediate contact with each other. Using this module the user can get the information about

Sandhi rules and processes. *Sutra* number in *Astyadhayi* and its description is displayed. User can learn three type of *Svara Sandhi*, *Vyanjan Sandhi*, *Hal Sandhi* through this *Sandhi* module Data is in Unicode. *Sandhi* exceptions and options are also incorporated.

This is the Flowchart for this module. These are the procedures written to make this module user friendly and more descriptive.



Input Procedure : This procedure is written to take the input from the user in the form of Unicode data. Phonetic keyboard is used for this purpose by which everybody can use this module easily.

CheckSandhi procedure: This procedure is used to find the *Sandhi* that appears with the given input.

Panini Ashtadhyayi Sutra Number: This procedure connects with Ashtadhyayi module. Click on Sutra number displays the sutras used.

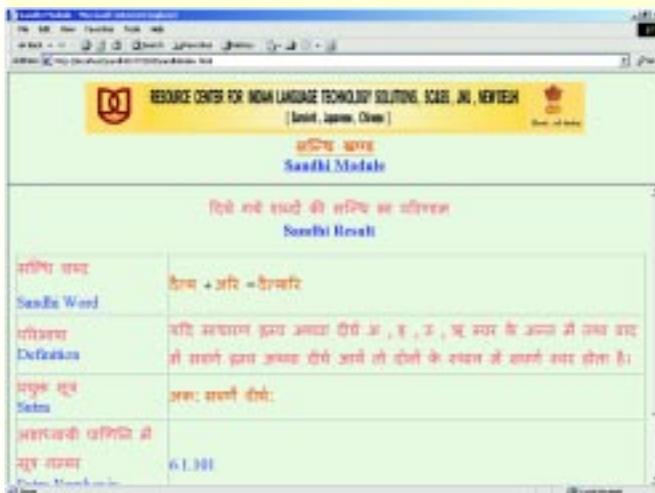
This module takes two words as input. First word cannot be null but second word can be. A user

can input the two words and submit the form to get the result of the given input.

- Fill the first word in the given textbox captioned 'Input First Word'.
- Fill the second word in the given textbox captioned 'Input Second Word'.
- Click the Submit Button.

Example- For the inputted first word "Dyata" and second word "ari" It shows all the explanations related to this *Sandhi* like its Sutra definition, which applies in process and the Sutra. It is also displays the Sutra number in *Ashtadhyayi*. This hyperlink leads to the more details of this sutra.

First input word + Second input word = resultant word



9. Japanese Lessons

The Japanese Lessons are very useful to introduce the Japanese Language. These are mainly for the Indians who know Hindi well can learn Japanese as the contents in the lessons are described in Hindi. Few Screen shots are given below.

Japanese Language contains a rich set of alphabets. It is important for the learner to be familiar with the set of alphabets. Keeping this in mind the most of fourteen lessons describe the basics of the Japanese Language with simple examples and explanations.

Grammars of the Japanese Language are discussed not elaborately but lightly to give idea of verb, tense, nouns, adjective etc. Care is taken not to make things complicated.

Kanji is another thing that adds a different value to the Japanese language. The language has a rich set of Kanji. Kanji's are explained with their stroke order.

Lots of facilities are presented to make the learning process easy and attractive. User can see all the words of the total lessons from any page. Again the new words of a section are also provided to the user. From these words list a user go to the Japanese-Hindi Lexicon for the details of the word. From the lessons of alphabet learning a user can go to the Japanese script learning module that see how the alphabets can be written and pronounced. These script learning includes all alphabets with Kanji.



10. Japanese Script learning and Character Recognition System

Scripts are the basic of a Language. And writing of these scripts are very important to learn a Language. Using this module the user learns Japanese alphabets and learns how to write and pronounce them. Writing strokes for a character are shown with stroke orders with four different speeds. It takes care of all characters with Kanji. Recognize a character after hearing its pronunciation is also over. It is also hyper-linked with Japanese Lessons. User can access this module from Japanese Lesson.

There are different types of alphabets in Japanese Language. All these are covered here. The writing process is made user friendly. There are four options to choose the speed of writing for an Alphabet. And there is audio support to listen the pronunciation of the alphabet. Emphasize is given on the stroke order

according to which the alphabets are correctly written.

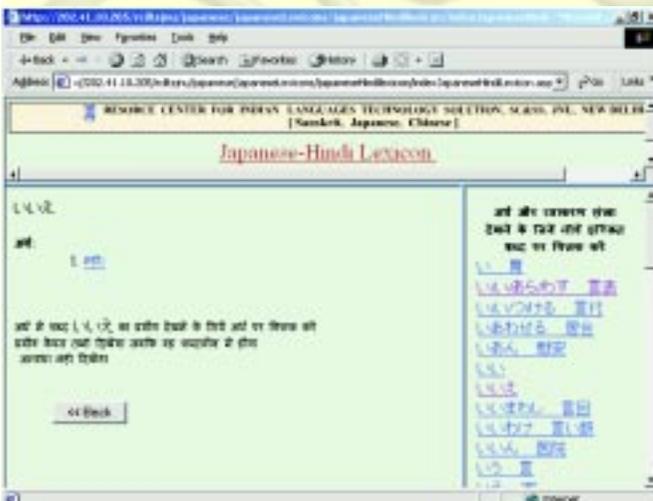
Except Basic alphabets there are numbers of Kanji in Japanese Language. Their formation and the writing process in the proper stroke orders are given.

Pronunciation of alphabet is important to speak a language. The attention is given for the right and clear pronunciation of the alphabets and audio supports are provided.



11. Japanese-Hindi Lexicon

In this module a user can get the attributes of a Japanese word in Hindi. It is very helpful to the user who know Hindi and wants to learn Japanese Language. The facility for getting input from the user is provided in both ways. Users can click the beginning letter of the word. It will display all words, which start with that letter in the right side window. On clicking particular word the meaning with phonetic notation and the grammatical category are displayed in the left window.



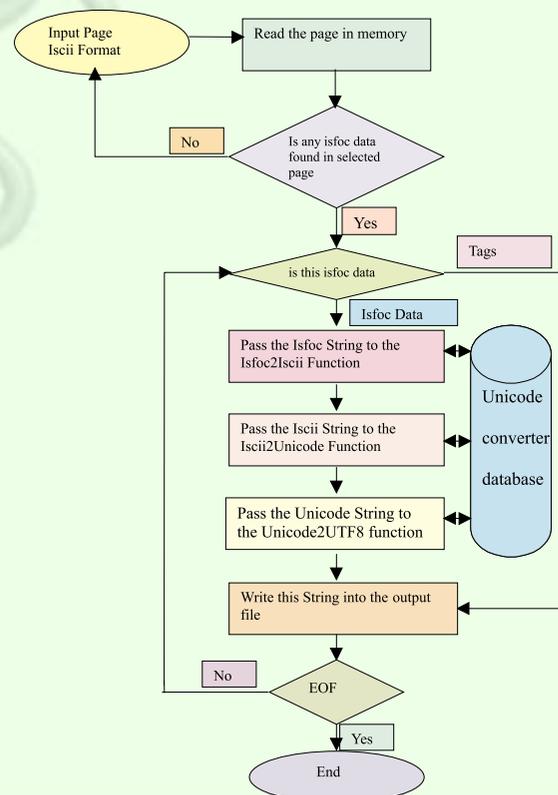
12. HTML Content Converter

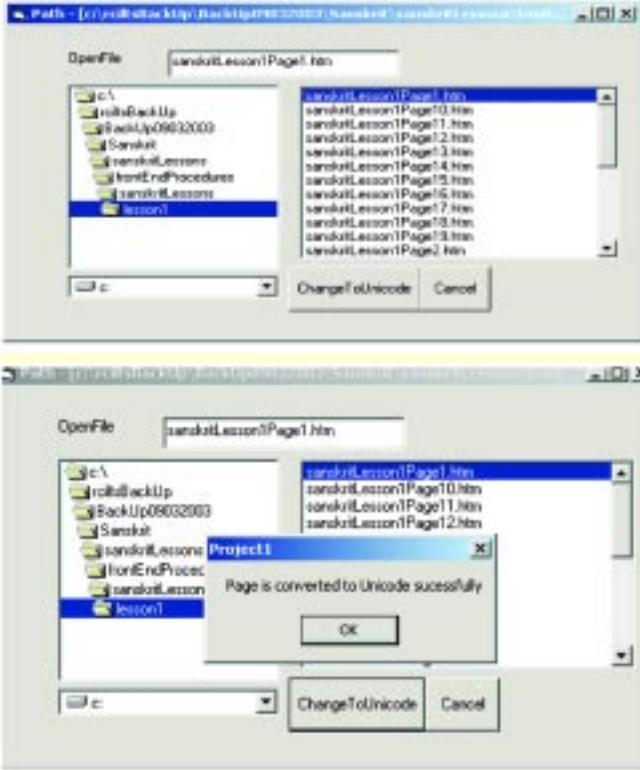
This converter converts ISCII/ISFOC based HTML content to Unicode based HTML content. It preserves the structure of the HTML content. It to a larger extent also preserves the presentation of the HTML content. It also incorporates the content encoding directive to UTF-8. So that the resultant output of the converter is complete in all respect.

Higher level design of the converter:

1. The converter reads the input file
2. Looks for the presence of ISFOC string inside the content
3. Whenever it finds an ISFOC string it converts them to ISCII
4. Thereafter it converts the ISCII string to Unicode data
5. It writes the resultant Unicode data to the output file.
6. It also updates the content encoding inside the output data

This is the flow chart of the HTML Content Converter.





Other Converters

1. ISCI Text to UNICODE Text Converter
2. Unicode Text to ISCI Text Converter
3. ISCI Database to UNICODE Database Converter
4. UNICODE Database to ISCI Database Converter
5. Devanagri_Unicode (Phonetic) Keyboard driver
6. Devanagri_Unicode (Inscript) Keyboard driver
7. Japanese_Unicode Keyboard drivers (under development)

13. Current/Future tasks

1. The utilities developed, being developed and to be developed at RCILTS, JNU will be available on the web.
2. RCILTS, JNU will participate in the development of at least two Devanagari Unicode Fonts (1 true type and 1 open type) locally, by procuring or by outsourcing.

3. The center will collect and make a repository of related work being done at national and international centers. Such as C-DAC, IIT Kanpur, IISC Bangalore, JETRO etc. The center will establish active contacts with these national and international bodies for acquisition of public domain material on the subject.
4. The format of the Lexicons will be advertised and discussed with other RCILTS
5. The center will collect and place on web some more texts such as Hitopdesh etc.
6. Center must enhance existing systems such as lexicons and developed systems such as Sandhi and Sandhi Vicched Modules leading to the development translation aid system for Sanskrit to other Indian Language and Voice-Versa. The centers must acquires Hindi/ Sanskrit Morphological Analyzers and then improve upon them.
7. Center will also enhance OCR performance for Sanskrit.
8. The Center will also enhance Spell Checkers, Basic Word Processors and Text Editors for Sanskrit.
9. The Center will procure a Message Server.
10. The task on OCR, Fonts, Spell Checker, Text Editor, Word Processor and Message Server will also be taken up. However, some of these activities may be taken up right away depending on the currently available funds with the resource center.

Following work being done at the center would be available fully to users on Internet ending September 30, 2003

Sanskrit Modules

1. Ashtadhyayi of Panini with Sidhant Kaumudi of Nagesh Bhatta (English Tika) and Prabhakari Tika (Hindi). In addition Katyayan's Vartik on Ashtadhyayi will also be available.

2. Dhatu Ratnakar with all forms of all Dhatus.
3. Pratyahar Module.
4. Ten Lessons on Sanskrit Sambhasn.
5. Samgya word forms (Pratipadic) with 1,000 words.
6. Sarvnam Module with all Pronouns.
7. Sanskrit-English Lexicon (Word Meanings only) with 30,000 words.
8. English-Sanskrit Lexicon (Word Meanings only) with 30,000 words.
9. 30 Sanskrit Lessons with Exercises.
10. Sandhi Prakaran with exceptions and options.
11. Lists of other Sanskrit words, such as adjectives, conjunctions etc.
12. Devanagri Script Learning Module with Matras and Conjuncts.

Japanese Modules

1. 21 lessons with exercises. These lessons shell constitutes a course which is more than equivalent to one year diploma course in Japanese.
2. A simple Hindi to Japanese translation exercise module.
3. Japanese-Hindi Lexicon (Word Meanings only) with 6,000 words.
4. Hindi-Japanese Lexicon (Word Meanings only) with 6,000 words.
5. Hiragana Script Learning Module.
6. Katakana Script Learning Module.
7. Kanjhi Script Learning Module.

In addition to adding more lessons the center plans to augment the lexicons with audio visual support, usage of words for meanings and add phrasal grammatical categories of words. Sandhi Vicched Module will be developed which will aid the development of simple translation systems

(Sanskrit to other Indian Languages and voice-versa). A simple Hindi to Japanese and Japanese to Hindi translators is also planned. The Center also plans to develop contents on Hitopdesh kind of material both for Sanskrit and Japanese.

There is non-availability of good expertise in the field of Chinese language. Centre is therefore of the opinion that the center may stop its efforts in developing systems for Chinese language and concentrate its energy on Sanskrit and Japanese.

*Courtesy: Prof. G.V. Singh
Jawaharlal Nehru University
School of Computer and Systems Sciences
New Mehrauli Road
New Delhi – 110 067
(RCILTS for Sanskrit, Chinese & Japanese)
Tel: 00-91-11-26107676,26101885
E-mail: gvs10@hotmail.com*