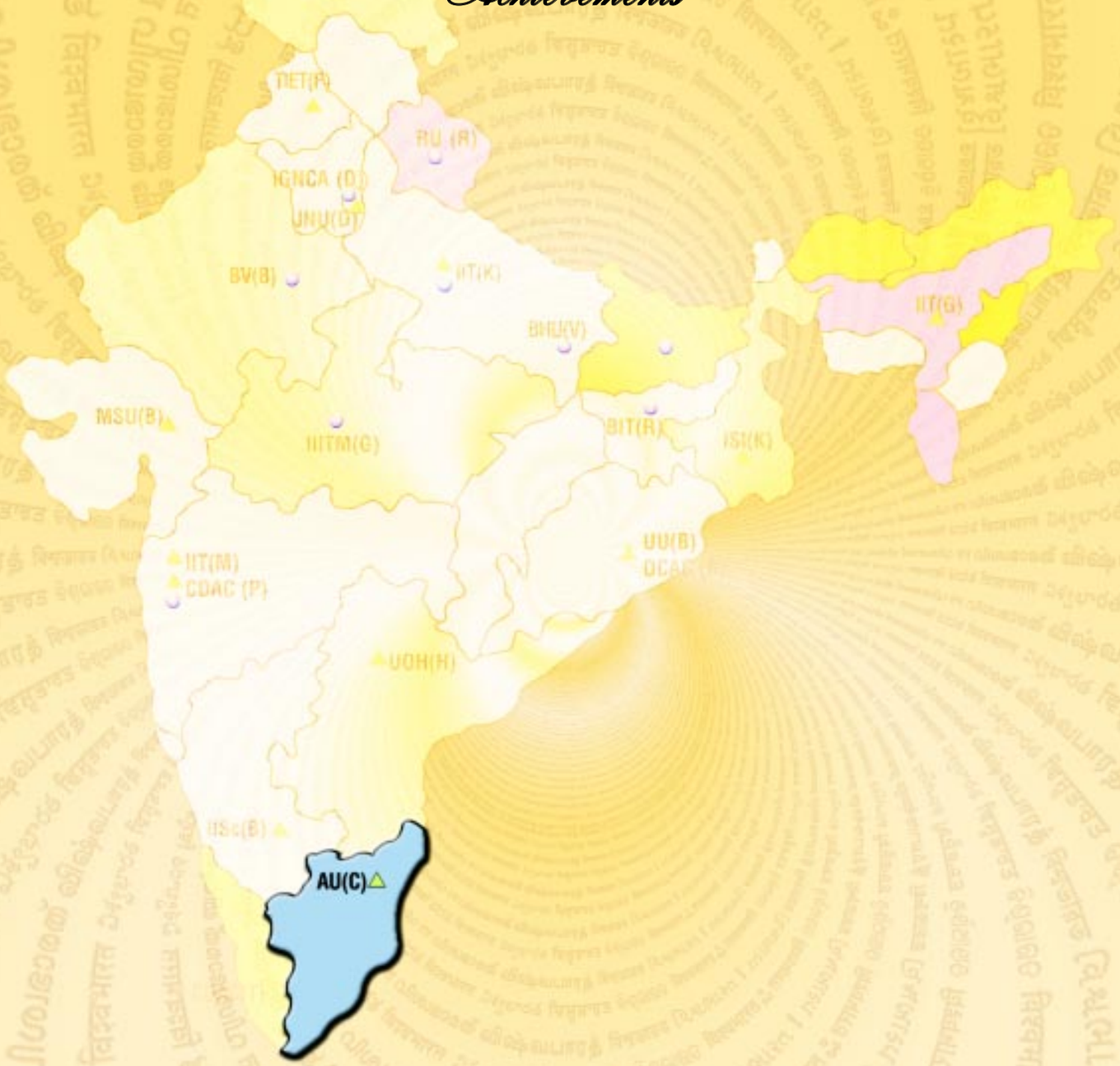
Resource Centre For
Indian Language Technology Solutions – Tamil
Anna University, Chennai

Achievements



School of Computer Science and Engineering
Anna University, Chennai-600025 India
Tel. : 00-91-44-2351265 Extn : 3340
E-mail : rctamil@annauniv.edu
Web Site: http://annauniv.edu/rctamil

RCILTS-Tamil
Anna University, Chennai

Introduction

The Resource Center for Indian Language Technology Solutions – Tamil, Anna University has been actively working in the area of language technology with special emphasis on the following areas:

- **Linguistic Tools**
- **Language Technology Products**
- **Content Development**
- **Research On Information And Knowledge Extraction**

Our website http://annauniv.edu/rctamil showcases the highlights of these activities. Downloadable demos are available for selected products.

A detailed report on the activities of the center is presented below.

1. Knowledge Resources

Screen shots of Knowledge Resources

The following activities have been undertaken for development of knowledge resources.

- **Online Dictionary**
- **Corpus Collection Tool**
- **Content Authoring Tool**
- **Tamil Picture Dictionary**
- **Flash Tutorial**
- **Handbook of Tamil**
- **Karpanaikkatchi - Scenes from Sangam literature**
- **District Information**
- **Educational resources**
  - **Appreciation of Tamil poetry**
  - **Geography for You**
  - **Chemistry for X Standard**

1.1 Online Dictionary

Online Dictionary is a language-oriented software for the retrieval of lexical entries from monolingual and bilingual dictionaries by many users simultaneously. It has been designed to interact with various language-processing tools – morphological analyzer, parser, machine translation etc.

The Tamil dictionary contains 20,000 root words. Each entry in the dictionary includes the Tamil root word, its English equivalent, different senses of the word, and the associated syntactic category. The root words are classified into 15 categories.

The special features of this Online Dictionary are the following :

- It is an inflectional or derivational dictionary. That is the search can be based on an inflected word. Given an inflected word, it calls the Morphological analyzer to get the root word and searches the dictionary. In addition to providing information about the root word

from the dictionary, it also provides the analyzed information about the inflection.

- It supports search based on syntactic category.
- It also supports search based on parts of a word.
- Tamil equivalent words of given English words can also be obtained.

Status : Dictionary is available with 20,000 root words. In effect, since it is an inflectional dictionary, it is possible to find the meaning of over 1 lakh words. It is to be enhanced to 35,000 root words.

1.2 Corpus Collection Tool

A corpus with 12 lakh words has been built from Tamil news paper articles, using a Corpus Collection tool. This tool automatically collects articles from different web-sites and stores them into a corpus database. The corpus database is organized into a word database and a document database. The databases have the following information : the article, category of the article, key words, title, date, author, source, unique words and their frequencies. The tool collects articles available in different fonts, and converts them to a standard font, before storing them in the database.

Status : Available for use with 12 lakh words. Collection is an on-going process.

1.3 Content Authoring Tool

This is a tool used to create and deliver content, integrating text, sound, video, and graphics. The design of the tool is generic so that it can be adapted to any domain. The information can be organized hierarchically with multiple levels. Content can be updated at any time, with new levels being added. There are primarily two different modes in which this operates – authoring mode, used by the content developer to organize and present the content, and the viewing mode used by the end-user to view the contents. The authoring mode stores the content in a database, and the viewing mode serves the pages using JSP. The user can select the sections that he/she wants to view.

Status : Testing in progress.

1.4 Tamil Picture Dictionary

This is a picture book for children. It has about three hundred words - organized as nouns and verbs. Each word has on associated picture, an explanation, the English equivalent and a small poem to illustrate its meaning.

Status : Available on our website with 300 words. Audio capability is to be added.

1.5 Flash Tutorial

This tutorial in Tamil, provides an easy way to learn Macromedia **Flash**, a content development tool. Flash is the most popular multimedia tool to enhance or design web pages or CBT packages. This tutorial contains a set of technical keywords along with the content. It starts with the basics of Flash and methodically moves up to an advanced level. The components available in the toolbox are explained. Creation of movies and animated images are also explained. The major concepts like Tweening and Symbol conversion are discussed in detail. Snapshots are provided wherever necessary. Plenty of examples are also provided. The English equivalent for the keywords are given within brackets.

Status : Available on our website

1.6 Handbook of Tamil

This e-handbook gives an overview and special features of Tamil grammar and literature. The handbook is divided into four sections - Introduction, Literature, Grammar and others. In the 'introduction' section, history of Tamil language and details about the Tamil speaking people are given. The second section provides an overview of Tamil literature starting from the Sangam age up to the modern times. The grammar section outlines the classification schemes used in Tamil grammar and highlights important features. The last section provides miscellaneous information such as names of days, months, measures, traditional art forms etc. in Tamil.

Status : Available on our website

1.7 Karpanaikkatchi - Scenes from Sangam Literature

This is a visualization of scenes from Sangam literature. Selected poems depicting various moods pertaining to Sangam period have been chosen. The poems, their interpretation and specially created imagery picturising the mood conveyed by these poems, have been presented.

Status: Development in progress

1.8 District Information

- Gives statistical, informative and special features about districts of Tamil Nadu
- User can browse through the information district wise or subject wise

Status : Under development

1.9 Educational Resources

- **Appreciation of Tamil Poetry**
- **Geography for You**
- **Chemistry for X Standard**

1.9.1 Appreciation of Tamil Poetry

The aim of this package is to introduce the nuances of Tamil poetry to young children through music, picture, animation and games. This package consists of 25 lessons that take the children from the basics of Tamil poetry to an advanced level of appreciation. Each lesson consists of the poem, segmentation of the compound words, word-by-word meaning, interpretation of the poem, related poems, and author information, and exercises.

1.9.2 Geography for You

This package is available both in Tamil and in English. The aim of this package is to systematically introduce the concepts of Geography, catering to students starting from primary to higher secondary stages. A total of 33 lessons have been organized accordingly. The topics covered include atmosphere, universe, solar system, earth, natural regions of the world, and the continents. Images, audio, and animation have been added to make this appealing to students.

1.9.3 Chemistry for X Standard

This package has been specifically targeted at tenth standard students. Here again the focus is on presenting concepts, and it is presented both in Tamil and English. The package consists of eight topics – periodic classification, atomic structure, chemical bonding, phosphorus, halogens, metals, organic chemistry, and chemical industry. In addition to innovatively using images, and audio, the highlight of this package is simulation of experiments using animation.

2. Knowledge Tools

The following tools have been developed for language and information processing.

- **Language Processing Tools**
  - **Morphological Analyser**
  - **Morphological Generator**
- **Text Editor**
- **Spell Checker**
- **Document Visualization Tool**
- **Utilities**
  - **Code Conversion Utility**
  - **Tamil Typing Utility**

2.1 Language Processing Tools

Tamil like many other Indian languages is a morphologically rich language. Formation of words occurs by means of agglutination. Thus each word consists of a root word and one or more suffixes. Hence morphological analysis of a word and morphological generation of a word are primary and challenging tasks. Hence, the core of the language processing tools developed are the **Morphological Analyser and Morphological Generator**. They are used to build other language processing tools and products, like parser,

translation system, spell checker, search engine, dictionary etc.

• Morphological Analyser

The morphological analyser tool takes a derived word as input and separates it into the root word and the corresponding suffixes. The function of each suffix is indicated. It has two major modules – noun analyser and verb analyser. A rule-based approach is used. The number of rules used is approximately 125. Heuristic rules are used to deal with ambiguities. A dictionary of 20,000 root words is used. The design of the analyser is such that the number of times the dictionary is accessed is minimized. The steps involved are as follows:

1. Given an input string, it starts scanning the string from right to left to look for suffixes. A list of suffixes is maintained.
2. It searches for the longest match in the suffix list.
3. It then removes the last suffix and determines its tag and adds it to the word's suffix list.
4. It checks the remaining part of the word in the dictionary and exits if the entry is found.
5. According to the identified suffix, it generates the next possible suffix list.
6. The process repeats from step 2, with the current suffix list.

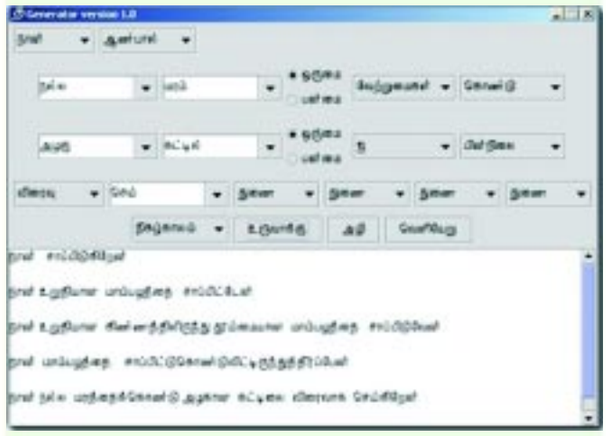
Screen shot of Morphological Analyser

Status : Currently, in the testing phase. A preliminary version with over 70% accuracy is available. It is being enhanced by adding rules, to improve the accuracy.

• Morphological Generator

The morphological generator does the inverse process of the morphological analyser. It generates words when Tamil morphs are given as input. It also has two major modules: - noun generator and verb generator. In the noun section, a root noun, plural marker, oblique form, case marker and postpositions are given as inputs. In the verb section a root verb, tense markers, relative participle suffix, verbal participle suffix, auxiliary verbs, number, person and gender marker are given as inputs. At a time, four auxiliary verbs can be added to the main verb. In addition to these, adjectives and adverbs are also handled.

It generates the word using the internal sandhi rules and the given inputs. It uses 125 morphological rules. For a given verb, with all possible combinations of tense markers and person number gender markers nearly 30 forms can be generated. With the auxiliary verbs added to the main verb, about 200 verb forms can be generated. In the same way for a given noun, with all possible combinations of case markers and postpositions nearly 68 forms can be generated.



Screen shot of Morphological Generator

Using this generator, an interactive sentence generator has been designed. Given the root noun and associative number, case, adjective,

postposition, etc., and the root verb and the corresponding auxiliary, tense, adverb, postposition, etc, this tool will generate morphologically and syntactically correct sentences. In order to generate a simple sentence at least a subject, verb and the tense should be given.

Status : Preliminary version is available. It is being enhanced by adding rules to take care of special cases.

2.2 Text Editor

The Tamil text editor is aimed at naïve Tamil users. It is platform-independent (works on both Windows and Linux machines). It provides basic facilities for word processing both in Tamil and English. It doesn't have any special file format. Files of both rich text format and text-only formats can be created or edited.

Status : Available for use. Can be downloaded from the web site.

2.3 Spell Checker

Tamil Spell Checker is a tool used to check the spelling of Tamil words. It provides possible suggestions for the wrong words. This spell checker is based on the dictionary, which contains Tamil root words. At present the size of the dictionary is 20,000 words.

The spell checker uses the morphological analyser for splitting the given Tamil word into the root word and a set of suffixes. If the word is fully split by the morphological analyser, it is assumed that the given word is correct. Otherwise it goes for the correction process. The following types of errors are handled :

1. Errors in case endings and PNG markers: In this case, if there are any errors in the case endings like il, aal, iTam, uTan, uTaiya, aan, aaL, arkaL, etc …, they are corrected and given to the suggestion generation phase.

2. Errors due to similar sounding characters, (eg. manam, maNam,palam, paLam, pazham, etc.) : In this case, if any letter in the erroneous word can have similar sounds, then it is replaced with equivalent letters and checked against the dictionary.

3. Adjacent key errors : Adjacent keys are based on Tamil 99 standard keyboard. In this case, each letter is replaced with adjacent keys and checked against the dictionary.

After the correction phase, all the corrected case endings, PNG markers, root words are given to the suggestion generation process. The spell checker makes use of the Morphological generator to generate all possible suggestions. User has the provision to select the suggestion among the list, ignore the suggestion or add the particular word to the dictionary.

Screen shot of Tamil Spell Checker

Status : Version 1.0 is available. Will be enhanced by increasing the root-word dictionary size to 35000.

2.4 Document Visualization

Document visualization aims at presenting abstracted information extracted from raw data files visually. Instead of forcing a user to read a whole document just to grasp certain information, document visualization offers very important clues about the document. It is a tool that is capable of visualizing any document.

The tool has various control points for the user to specify his options. The tool takes any tagged document as input and extracts the tag information and uses that for visualization. The various views presented are scatter plot view, to show relative tag positions in a two dimensional plane, document zoom view, to show thumb views of pages, statistical information view, to show statistics about input document and synchronous

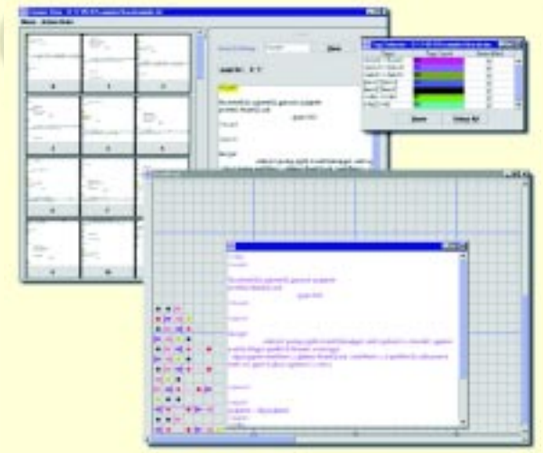
depth view. These views together make up the entire tool.

Scatter Plot view : Here the document is mapped on to a graph. This view gives the relevant position of the tag in a particular page. It normalizes the tag position according to the display. The position of the tags is marked with circles of predefined size and color. Their shapes distinguish nested and normal tags. Apart from displaying the tags, it gives the relevant information about a particular tag in a separate screen, when the user clicks on it and, if the selected tag is of nested type, it also drops a line from the starting position to the end of the tag.

Zoom view : The document zoom view prepares the intermediate file, which has a proper alignment of certain number of lines per page. Each page is shown as a small thumbnail. The user can click on the appropriate page to view that page. Search can be done in zoom view.

Statistical Information view : This view presents the frequencies of tags identified in the input file in a bar chart form. It is used to show the legend about the tags present in the document.

Synchronous Depth view : This view projects the input document onto the 3-dimensional plane. The whole document is considered to be an object in 3- dimensional plane. This view helps to get the picture of how the tags spread over the document.



Display of Zoom and Scatter Ploter View

Status : A demo version is available. Currently being tested.

2.5 Utilities

In order to facilitate language processing in Tamil, a set of common utilities have been developed. These include **Code conversion utility, and Tamil typing utility.**

Code Conversion Utility

There are a plethora of fonts in use on various web sites in Tamil. Hence a generic code conversion utility has been developed. Given a source and a target-coding scheme, this utility will generate the code for the conversion. This code can be integrated as part of any language processing tool or application. Existing standard coding schemes are already built in. Any other coding scheme has to be explicitly specified.

Status : Available for use.

- **Tamil Typing Utility**

This is an API, which indirectly serves as a keyboard driver. It allows the user to type the Tamil text in both transliterated mode and using the Tamilnet99 standard keyboard. This can be included as part of any language processing application that requires Tamil input.

Status : Available for use.

3. Translation Support Systems

The following are the components that have been developed to aid machine translation :

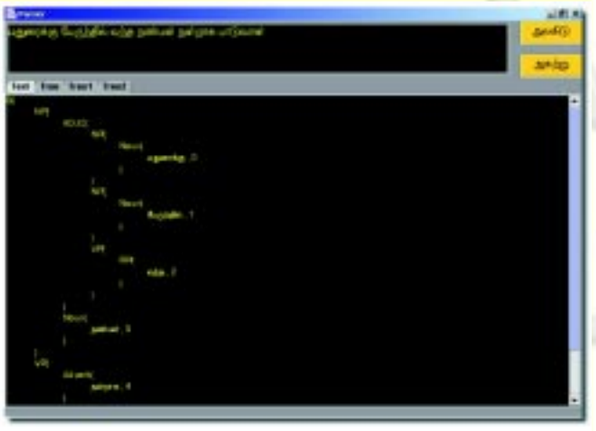
- **Tamil Parser**
- **Universal Networking Language (UNL) for Tamil**
- **Heuristic Rule Based Automatic Tagger**

3.1 Tamil Parser

The Tamil parser identifies the syntactic constituents of a Tamil sentence. It outputs the parse tree in a list form. The free word order feature of Tamil makes parsing a challenging task,

since there is a need to associate and link syntactic components that are not always adjacent to each other. This is done by associating positional information with the words. The parser tackles both simple sentences having a verb, a number of noun phrases, and simple adverbs and adjectives, and complex sentences with multiple adjectival, adverbial, and noun clausal forms.

The processing uses a phrase-structure grammar together with a limited amount of look-ahead to handle free word order. The words of the sentence are first analyzed using the morphological analyzer, and the root word along with the part-of-speech tag and suffix information obtained is input to the parser. However, the analyzer does not unambiguously give the part-of-speech information for all words. In these cases, linguistic based heuristic rules are used to obtain the associated tags. Currently, 15 heuristic rules are used. For sentences with multiple clauses, the parser first identifies the cue words of the clauses and syntactically groups the clauses based on the cue words and phrases.



Screen shot of Tamil Parser

Status : At present it can handle simple sentences, and complex sentences with a noun clause, with multiple adjective clauses, with multiple adverb clauses, and both multiple adjective and adverb clauses.

3.2 Universal Networking Language (UNL) for Tamil

Universal Networking Language (UNL) is an intermediate representation for inter-lingua based machine translation. The Universal Networking Language (UNL) is an intermediate language based on semantic network, and used by machines to express and exchange natural language information.

UNL provides sentence-by-sentence representation of the semantic content. Sentence information is represented as a hyper-graph having Universal Words (UWs) as nodes and binary relations as arcs. Binary relations are the building blocks of UNL sentences. They are made up of a relation and two UWs. Each relation is labeled with one of the possible label descriptors. Relations that link UWs are labeled with semantic roles of the type such as *agent, object, experiencer, time, place, cause*, which characterize the relationships between the concepts participating in the events. This project focuses on developing the Tamil EnConverter (Tamil to UNL format) and Tamil DeConverter (UNL format to Tamil). Both the EnConverter and the DeConverter use language dependent rules, and a language dictionary that contains Universal Word (UW) for identifying concepts, word heading and syntactical behavior of words.

An EnConverter is a software that automatically or interactively enconverts natural language text into UNL. EnConverter is a language dependent parser that provides synchronously a framework for morphological, syntactic and semantic analysis. EnConverter uses the following sequence of operations:

- Convert or load the rules
- Input a Tamil sentence
- Apply the rules and retrieve the word dictionary
- Output the UNL expressions

EnConverter analyses a sentence using the word dictionary, knowledge base, and enconversion rules. It first uses the morphological analyser to analyze the given word. It retrieves relevant dictionary entries from the language dictionary, operates on nodes in the Node-list by applying

enconversion rules, and generates semantic networks of UNL by consulting the knowledge base.

A DeConverter is software that automatically or interactively deconverts UNL into natural language text. Tamil DeConverter uses the morphological generator to generate the Tamil sentence.

The information needed to construct the UNL structure and generate Tamil sentence from UNL structure is available at different linguistic levels. Tamil being a morphologically rich language allows a large amount of information including syntactic categorization, and thematic case relation to be extracted from the morphological level. Endings in Tamil words with syntactic functional grouping and semantic information are used to identify the correct binary relation. For syntactical functional grouping, information about relating concepts like verbs to thematic cases, adjectival components to nouns and adverbial components to verbs are needed. Syntactical functional grouping, has been done by the specially designed parser, taking into consideration, the requirements of the UNL structure.

Status : At present, all the relations have been identified and simple sentences can be handled.

3.3 Heuristic Rule Based Automatic Tagger

This tagger automatically tags the words of a given document with parts of speech tags. It uses linguistically oriented heuristic rules for tagging. It does not use the dictionary and the Morphological Analyser.

The part of speech (POS) tagger uses heuristic rules to find the nouns, verbs, adjectives, adverbs and postpositions. This is done, by checking the appropriate morphemes of the words. Clitics are removed from the word before it goes through rule verification.

Some of the heuristic rules used are as follows.

1. If a word contains PNG markers and tense markers, then it is tagged as verb.
2. If a word contains case marker then it is tagged as noun.
3. Standalone words : Some commonly used words are categorized and stored in lists. If a word belongs to a particular list then the word is tagged with that category.
4. Fill in rule: If the unknown word comes in between two nouns, then it is tagged as a noun.
5. Verb Terminating: If the unknown word comes at the end of a sentence, then it is tagged as a verb. This is due to the fact that normally Tamil sentences end with a verb.
6. Bigram : The unknown word is identified using the category of the previous word.

Status : About 83% accuracy is obtained when tested on a sample corpus.

4. Human Machine Interface Systems

Two packages that serve as human-machine interfaces, have been developed. They are :

- **Text To Speech (Ethiroli)**
- **Poonguzhali (A chatterbot)**

4.1 Text To Speech (Ethiroli)

Ethiroli is a Text to Speech engine for Tamil. The engine has been so designed that it can be plugged into any application requiring a text to speech component. When Tamil text is given as input, it is pre-processed, and transliterated using linguistic rules, to a phonetic representation in order to remove homographic ambiguities. Homographic ambiguity arises in Tamil since the same character can have different sounds depending on the position in which it occurs. For example, the vallinum letters (è, ê, ì, î, ð, ø) in Tamil have more than one sound based on their position of occurrence within a word. Each word in this transliterated text is then split into its corresponding phonemes using the cvcc model. Once the phonemes are identified, the sound files corresponding to the phonemes are concatenated and played in sequence to give the speech output.

Currently, 4000 syllables are used to generate Tamil speech using this method.

Status : A preliminary version is available. Efforts are on to improve the quality of speech produced.

4.2 Poonguzhali (Chatterbot)

Poonguzhali is a generic Tamil chatterbot that allows the user to chat with the system on any technical topic. A chatterbot is an AI program that simulates human conversation and allows for Natural Language communication between man and machine.

A question or a statement is taken as input from the user. The function of the system is to generate an appropriate response, based on the context of the input. The user can choose any existing topic for conversation and ask questions to the system in Tamil. Provision for input in transliterated form is available. The system identifies the minimal context of the input as to what the user is trying to ask, no matter in what way the user frames his question. This is done using a set of decomposition rules. The response is then formed using a set of reassembly rules that reside in the knowledge base. This response is then reframed to match the way in which the user had framed his question. The system is also versatile enough to initiate a conversation based on the earlier sequence of the conversation, when there is no input from the user.

The software offers the facility of an optional alternate explanation to the previous one, in case the user doesn't understand the earlier answer given by the system. The pronoun references like the pronoun "idhu" in the second question related to the first one are identified. The Yes/No type of questions and compound technical terms are also handled.

Currently there are two domains available. Any number of domains can be added. For this purpose, the tool provides a separate interface for adding domains and for updating the knowledge base. This has three separate options - for entering the knowledge base, non-technical words and technical terms. The input key is entered with its response and alternate response in the knowledge base. The input key can be more than one word with the '+' separator.

Screen shot of Poonguzhali (Chatterbot)

Status : A version with knowledge base for two domains is available.

5. Localization

A complete office suite for processing Tamil information has been developed. This works on both Windows and Linux platforms. A search engine to search Tamil websites has also been developed.

- **Tamil Office Suite**
  - **Tamil Word Processor,**
  - **Presentation Tool for Tamil,**
  - **Tamil Database**
  - **Tamil Spreadsheet**
- **Tamil Search Engine**

5.1 Tamil Office Suite (Aluval Pezhai)

The Tamil office suite consists of four text processing applications, namely :

- Tamil word processor

- Presentation tool for Tamil
- Tamil Database
- Tamil Spreadsheet.

This office suite is specifically designed for Tamil users. However, it can be easily localized for any language.

• Tamil Word Processor (Palagai)

The Tamil word processor is aimed at Tamil users. It provides basic facilities for word processing both in Tamil and English. There are two versions of the word processor available. One with the spell-checker and grammar-checker, and another without the spell checker and grammar-checker. Both these versions include an e-mail facility. The grammar-checker checks for person, gender and number compatibility between the subject and the main verb of the sentence.

The word processor does not have any special file format. Files of both rich text format and text-only formats can be created or edited. HTML files can be viewed with this word processor. E-mails can be sent using this word processor if the current system is connected with the SMTP server. Email attachments are also handled.

Status :Version 1.0 is available for use.

• Tamil Presentation Tool (Arangam)

Arangam is a presentation tool for Tamil. It organizes a presentation with slides consisting of text, pictures and images. The presentation is saved in a proprietary format.

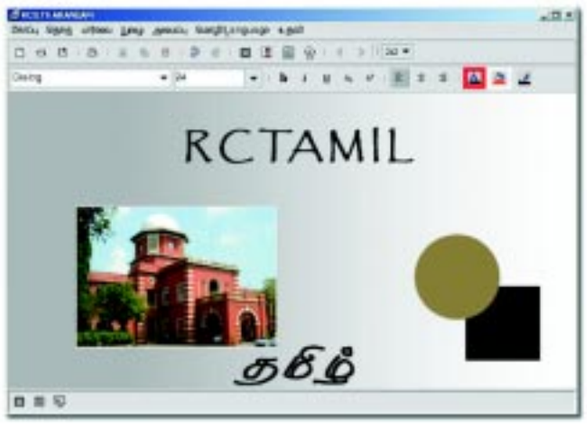
Once a file is created one can add, remove or edit slides with common operations like adding text boxes, pictures and general shapes. There are three modes of operation:

Edit Mode (Slide)

Preview Mode (Slide Sorter)

Presentation Mode (Slide Show)

The edit mode is the one that allows working on one selected slide. Preview mode gives a preview of number of slides (there are different possibilities 2x2, 3x3 or 4x4 slides in one page). Presentation mode is the regular slideshow.



Screen shot of Tamil Presentation Tool (Arangam)

Status : Version 1.0 is available for use.

• Tamil Database (ThaenKoodu)

The database tool helps a user to store Tamil data and also provides various means of retrieving it using queries, forms and reports. The help document for this tool is provided in bilingual format (Both Tamil & English). All the data are stored in MS-Access format.

This tool is divided into five major modules. They are Database, Table, Query, Forms and Reports modules. In the Database module, a new database can be created or existing databases can be edited.

The Table module allows the user to create the tables in two different ways. One is the design mode where the table name, field name, data type and the constraints (null, primary key and so on.) are to be entered. The other is through the Table creation wizard, where the sample table names along with their fields are provided to create the table. Tables can be modified and the structure of the table can also be altered. New fields can be added to the existing table at any time.

The Query module handles queries in three different modes. In the first, the table name and field name should be entered to query the table. The second mode uses the wizard, where the

required table name and the corresponding field name can be selected. The third one is, where the Tamil query can be entered in the space provided, to retrieve the result.

The Forms module generates forms. The forms can be built with the help of a wizard and can also be saved for later use. Report module allows the user to generate reports. The report can also be saved and printed.

Screen shot of Tamil Database (ThaenKoodu)

Status : Version 1.0 is available for use.

• **Tamil SpreadSheet (Chathurangam)**

Tamil Spreadsheet allows one to easily enter or edit the data in the desired format, save data, and view using various charts. Mathematical expressions are also handled. The data is saved in the proprietary format (".tss" Tamil spread sheet format).

It consists of three modules viz. Worksheet module, Expression Module and Chart Module. The default number of worksheets available is three. More number of worksheets can be added. The worksheets can be deleted or renamed. Worksheet module allows the user to enter the data in the required format. The basic operations like cut, copy, paste, find, replace, delete sheet, print the sheet, insert row/column and delete row/column are available in this tool.

The Expression Module evaluates the expressions. Addition (Sum), Count, Maximum, Minimum and Average are the functions supported in this module.

The Chart Module allows the user to draw the charts, which are visually appealing and make it easy for users to see comparisons, patterns, and trends in data. This module represents the data in the form of pie, horizontal bar and vertical bar charts. It can be saved as jpeg (.jpg) file format. The chart title and the scale to draw the charts can also be customized.

Status : Version 1.0 is available for use.

5.2 Tamil Search Engine (Bavani)

A search engine is a software package that searches for documents in the Internet dealing with a particular topic. The Tamil Search Engine aims at looking up Tamil sites for information sought by a user. It searches for Tamil words in Tamil web sites available in popular font encoding schemes. Currently 22 font-encoding schemes are supported. An important feature of the search engine is the integration of the morphological analyser, so that the keywords are always reduced to the root form.

The system gathers information from the Internet by the process of crawling. The information is gathered and stored in the database. An interface is provided to the user to enter the query. The query is analyzed with the information from the database, and the information that matches the query is returned to the user.

The following are the important modules handled in the search engine:

Crawler: This is the kernel part of the design. Starting from a given URL and domain, the document will be downloaded and parsed. It will then retrieve the URLs from that document and insert these URLs in the URL tree. After parsing each document, the next URL in the tree will be taken out for further crawling.

Database: A database will be maintained to store each word and its corresponding appearance in all the documents. These words are stored as a B-tree for efficient indexing. The position of the

word in various HTML tags is maintained and used for ranking analysis.

Searcher: Searcher accepts queries from the user, and then splits the queries into words. The given words are searched in the BTree and finally the results are sent back to the user after ranking is done. The ranking is performed in the order of number of outgoing links, number of occurrences, description (meta), head and title. The results are displayed in the order of AND, OR operations. The document, which has the maximum number of query words, gets a higher rank.

Status : Available for use.

6. Language Technology Human Resource Development

Many activities were undertaken to create awareness and develop human resources in the area of Language Technology. The activities can be grouped as follows:

- **Creating awareness among the general public**
- **Language Technology Training**
- **Co-ordination with linguistic and subject experts**
- **Co-ordination with Government and industry**

6.1 Creating Awareness Among The General Public

One of the priorities of the center has been to create awareness about the need for content development among the public. Working towards this goal, we targeted school and college teachers and students and the following workshops have been conducted:

- Workshop on Multimedia Tools for Effective Teaching: Conducted a two day workshop on Multimedia Tools for effective Teaching for college and school teachers in Feb 2001 – 100 teachers attended the workshop (200 person days).
- Workshop on Multimedia Tools for content creators Conducted a ten-day hands-on workshop on Multimedia tools for content creators in May 2001 – 13 teachers attended the workshop (130 person days)
- Effective Teaching through computer

  A presentation was given by RCILTS-Tamil members at S.B.O.A School, Anna Nagar, Chennai. 20 teachers attended.

  A presentation was given by RCILTS-Tamil members, at Teachers training school, Saidapet, Chennai. 100 teacher trainees attended.
- WWW-2001

  Organized a Creative writing competition for school children on Indian cultural aspects in Tamil and English, titled " WWW 2001, Weave your Words to the Web", in Sep 2001 to create an awareness about content development. 40 students participated.
- InSight

  "InSight – Scientific Thought, the Indian way", an Indian sciences appreciation course for school children, is being conducted every 6 months. (500 person days)

6.2 Language Technology Training

Language technology training has been imparted in two ways. One is by conducting intensive workshops to encourage research and development in language technology, and the other is by carrying out development projects in-house.

- Workshop conducted:
  - Organized a workshop titled "Corpus-based Natural Language Processing" targeted at both computer scientists and linguists
    - Pre workshop Tutorials – Dec 14 –16, 2001
    - Date: Workshop Dec 17, 2001 – Jan 2, 2002
    - No. of participants: 40

- Contents: Paninian Grammar, Dependency Grammar, Machine Translation, Statistical NLP, TAG grammar etc.

• In-house Training

More than 15 project associates have been actively working in the area of language technology for the past three years and have gained expertise in this area. Specifically they are well versed in computational morphology, syntax, lexical formats, and font and coding issues.

A lot of enthusiasm and interest has been created among students to work in the area of language technology. Over the last three years about 30 major student projects have been carried out in language technology. The students are both from under graduate (B.E – 36 students – 12 batches) and post graduate (M.E and M.C.A – 18 students) streams. Some of these projects have been further developed into products.

A number of research scholars are actively persuing research in this area. Three M.S by research scholars are working in the areas of language translation, speech processing, and information retrieval. Three Ph.D scholars are working in the areas of information extraction, Indian logic based knowledge representation, and automatic software document generation.

6.3 Co-ordination with Linguistic and Subject Experts

The center has been coordinating with various experts and agencies both for linguistic expertise and for content development. Numerous interactions have been organized for this purpose.

- Seminar conducted by RCILTS-Tamil members with S.B.O.A School, Anna Nagar, Chennai, for content development in Tamil Literature.
- Seminar conducted by RCILTS-Tamil members with MOP Vishnawa College, Chennai, for content development in Tamil Culture.
- Discussion with Tamil scholars for effective content development in classical Tamil literature.
- Seminar given by Dr.Subaiya Pillai and Dr.D.Ranganathan at Anna University for "Sandhi in Tamil".
- A one day workshop titled "Language and Computational Issues" for interaction with Tamil Linguists

  Date: Feb 15, 2002

  No. of participants: 30
- Discussion with experts from C-DAC, ASR, Melkote, CIIL Mysore, Annamalai University, Madras University regarding language processing tools developed.
- Demonstration of products developed at various forums.

6.4 Co-ordination with Government and Industry

The products developed have been demonstrated to industries working in the area of language technology. The outcome of this interaction is two-fold. One is the signing of memorandum of understanding with industries like Modular Infotech, Pune, Apple Soft, Bangalore, and Chennai Kavigal, Chennai for products such as spell checker, morphological analyser and generator, and text to speech engine. The other is the proactive interest shown by Kanithamizh Changam, a consortium of industries working on Tamil computing to work in close coordination with the center.

The office suite will be made available free of cost to interested users, industries and Government organizations. It is proposed to publicize the availability of the products and tools through interaction with appropriate state government authorities. Core technology such as morphological analyser, generator, and spell checker will be made available after signing of appropriate MOUs. The educational resources developed will be made available freely to government schools.

7. Standardization

All the software developed caters to both the prevailing Tamil Nadu Govt. standard (TAM/TAB font encoding scheme), and the ISCII standard. Testing for Unicode support is in progress. The center, actively co-ordinates with the Tamil Nadu Govt, in defining the Tamil Unicode scheme. Available standards for transliteration have been adhered to, wherever necessary.

All the development work has been carried out in Java, to provide cross-platform operability. All the software has been tested on both Windows and Linux platforms. Code, technical and user documentation have been done for all the products.

8. Research Activities

The following research activities have been going on, in the areas of information retrieval from documents and knowledge representation.

- **Text Summarization**
- **Latent Semantic Indexing**
- **Knowledge Representation System Based On Nyaya Shastra**

8.1 Text Summarization

Document summarization plays a vital role in the use and management of information dissemination. This project investigates a method for the production of summaries from arbitrary Tamil text. The primary goal is to create an effective and efficient tool that is able to summarize large Tamil documents. Summarizer proceeds to produce a meaningful extract of the original text document using the sentence extraction model. Lexical chains are employed to extract the significant sentences from the text without complete semantic analysis of the original document. Segmentation is done initially which helps to maintain the coherence in the generated summary.

The three important steps involved in summarization are : Segmentation of the document, Lexical chain creation and significant sentence extraction for summary generation. Tamil WordNet is used as knowledge source for summarization, which helps to find out the relations to build lexical chains. It is able to incorporate the relationship among words and at the same time account for the various senses of the words.

Segmentation: Text segmentation is a method for partitioning full-length text documents into multi-paragraph units that correspond to a sequence of sub topical passages. This method is based on Multi-paragraph segmentation of expository text. It has now been implemented for Tamil.

Lexical Chain Creation : Lexical chains are formed by disambiguating the senses of the words occurring in the original text and chaining the related terms together. Though repetition of words itself carries informational value, collocation of terms serves to recognize topics precisely. Ranking of the lexical chains and selecting the strong chains are done using Length and Homogeneity index (the number of distinct occurrences divided by the length.).

Sentence Extraction : Once the chains have been selected, the next step of the summarization algorithm is to extract full sentences from the original text based on chain distribution. There are methods, which involve choosing the sentence that contains the first appearance of a chain member in the text. Depending on the requested summary ratio, this algorithm can dynamically change the block size and token-sequence size in the segmentation step so that the number of segments recognized is either decreased or increased. The number of segments in turn determines the number of lexical chains thus either increasing or decreasing the length of the resulting summary. For convenience, the percentage size of the document is represented in terms of the number of sentences.

Status : A demo version for sports domain is available.

8.2 Latent Semantic Indexing

Latent Semantic Analysis is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. Latent Semantic Indexing (LSI) is chosen to index the documents, for it uses the higher order associations between the words in a passage or document. This information is further represented mathematically for easier manipulation. Tamil language follows free word order, and Latent Semantic Indexing does not consider word order to retrieve the semantic space of the documents. Hence, the choice of LSI for indexing Tamil documents.

In LSI, the raw text is represented as a matrix in which each row stands for a unique word and each column stands for a text passage or other context. The Matrix plays a vital role in LSI. It is considered to be the semantic space. Each cell contains the frequency with which the word of its row appears in the passage. Singular value decomposition is applied to the matrix.

The following applications have been done using the LSI techniques:

Sentence Comparison : The similarity of multiple sentences in Tamil is assessed. A similarity score for each submitted sentence , will be computed and given to the user.

Essay Assessor : This application compares two essays. The essays are accepted as input, separated, and converted as vectors. Then the similarity between them is computed, and given as a score to the user.

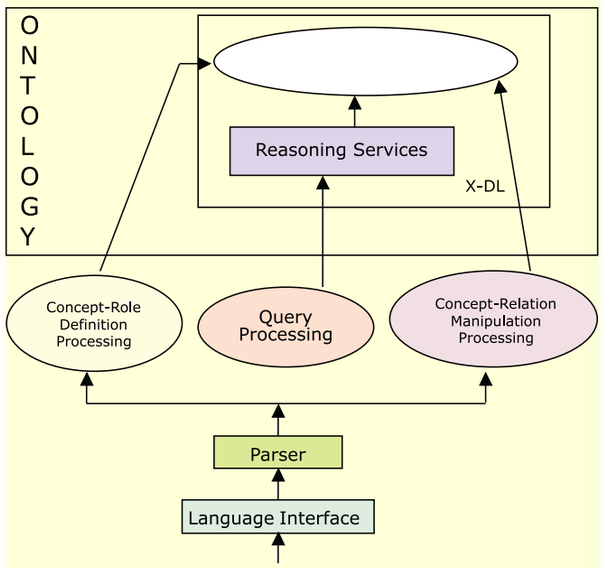
Document Retrieval: Relevant documents are retrieved using Latent Semantic indexing. In this application, the query words are the input , the relevant document is retrieved based on the query. The result is ranked and presented to the user.

Status : A demo version is available.

8.3 Knowledge Representation System Based On Nyaya Shastra

World knowledge representation for the purpose of language understanding is a challenging issue because of the amount of knowledge to be considered and the ease and efficiency of retrieval of appropriate knowledge. An important issue in the design of an ontology dealing with world knowledge is the philosophy on which the classification scheme is based. In this work, an ontology based on Nyaya shastra, an Indian Logic System has been designed and a knowledge representation system called KRIL has been implemented. The methodology is based on the notion of associating qualities and values to concepts and adding different kinds of negation, which brings a new perspective to the interpretation of knowledge. The knowledge hierarchy and the techniques to interpret the knowledge have been adapted from Nyaya shastra. Association of qualities and values to concepts and addition of new relationships based on these associations and new relations between concepts themselves adds a new dimension to the reasoning process.

To adapt the classification scheme and reasoning methodology of Nyaya shastra, description logic has been extended by adding operators for concept definition with qualities and for new types of negation. Hence, the building blocks of this model are the concept, inherent association of qualities, values, relationship between concepts and relationships between concept and qualities. The representation of knowledge at various levels includes time variant and invariant components. The special feature of this model is handling of prior and posterior absence of concepts through additional negation operators. This model can be used as a base for various applications like natural language processing, natural language understanding, machine translation etc.



Screen shot of KRIL-system

Status : An implementation of this model to represent the knowledge about the dairy domain has been done.

9. Publications

1. G.V.Uma, and T.V.Geetha (2001), "Generation of natural language text using perspective descriptors in frames", IETE Journal of Research-special issue on Knowledge and Data Engineering, January-April 2001, Vol.47 No.1&2 pp43-56.

2. G.V.Uma, and T.V.Geetha, "Softplan a planner for automatic software documentation", National Conference on Document analysis and recognition, Mandya, 2001, pp 266-271.

3. G.V.Uma, T.V.Geetha, "Automatic software documentation using frames and causal link representation" VIVEK journal – A quarterly in Artificial Intelligence, 2001, vol.14, No.3 pp 3-13.

4. D.Manjula and T.V.Geetha, "Message Optimization by Polling for Text Mining", National Conference on Document analysis and recognition, Mandya, 13-14 July 2002, pp. 199-202.

5. D.Manjula, P.Malliga and T.V.Geetha, "Semantic based Text Mining", First International Conference on Global Word net, Mysore, 21-25 January 2002, pp. 266-270.

6. D.Manjula, and T.V.Geetha, " Distributed semantic based text mining", To appear in the proceedings of the Third International Conference on Data mining methods and databases for engineering, finance and other fields, Bologona, Italy, 25-27 Sep 2002.

7. G.Aghila, Ranjani Parthasarathi and T.V.Geetha, "Design of conceptual ontology based on Nyaya theory and Tolkappiam", Third International conference on South Asian Languages (ICOSAL – 3), Jan 4 – 6 University of Hyderabad, Hyderabad, 2000, pp.2.

8. G.Aghila, Ranjani Parthasarathi and T.V.Geetha, "Indian Logic Based Conceptual Ontology Using Description Logics", National Conference on Data Analysis and Recognition, Mandya, 2001, pp.279-286.

9. Devi Poonguzhali, P.Kavitha Noel, N.Preeda Lakshmi, T.V.Geetha, A. Manavazhan, "Tamil Wordnet", First International Conference on Global Wordnet, 21-25 Jan 2002, pp 65-71.

10. G.V.Uma, and T.V.Geetha, "A knowledge based approach Towards automatic software documentation", Spring 2002, International Journal of Trends in Software Engineering Process Management, Malaysia.

11. G.V.Uma, N.S.M.Kanimozhi, and T.V.Geetha, "Automatic software documentation in Tamil", presented in the 3rd International conference on south Asian languages ICOSAL-3", University of Hyderabad, 1999.

12. Madhan Karky.V, Sudarshanan.S, Thayagarajan.R., T.V.Geetha, Ranjani Parthasarathy, and Manoj Annadurai, "Tamil Voice Engine", Presented at Tamil Inayam, Malaysia, 2001.

13. P.Anandan, T.V.Geetha, and Ranjani Parathasarathy, "Morphological Generator for Tamil" Presented at Tamil Inayam, Malaysia, 2001.

14. G.V.Uma, and T.V.Geetha, "A knowledge based approach Towards automatic software documentation", Spring 2002, International Journal of Trends in Software Engineering Process Management, Malaysia.

15. G.V.Uma, N.S.M.Kanimozhi, and T.V.Geetha, "Automatic software documentation in Tamil", presented in the 3rd International conference on south Asian languages ICOSAL-3", University of Hyderabad, 1999.

16. D. Manjula and T.V. Geetha, " Semantic based text mining", of the International conference on Global WordNet, Pg 266-270, Mysore, Jan 21-25, 2002.

17. D. Manjula, A. Kannan and T.V. Geetha "Semantic Information Extraction and Query processing from world wide web", Bombay, Dec18-21, KBCS 2002.

18. D. Manjula and T.V. Geetha, "Message Optimization using polling for distributed text mining," National Conference on Analysis and Recognition, Mandya, 2001.

19. P. Malliga, D.Manjula and T.V. Geetha, " Boostexter for Tamil Document Categorization", International Conference on South Asian Languages .

20. Siva Gurusamy, D. Manjula and T.V. Geetha, "Text Mining in request for comments Document Series", LEC 2002 Language Engineering Conference, Dec 13-15, Hyderabad, 2002.

21. T.Dhanabalan, T.V.Geetha, " UNL – EnConverter for Tamil", International Conference on South Asian Languages (ICOSAL) – 4, Dec 3-5, Annamalai University, TamilNadu, 2002.

22. T.Dhanabalan, K.Saravanan, T.V.Geetha, "Tamil to UNL EnConverter" International Conference on Universal Knowledge and Language (ICUKL) – Nov 25-29th, Goa, India.

23. G.S.Mahalakshmi, G.Aghila, T.V.Geetha, "Multi-level Ontology representation based on Indian Logic system" International Conference on South Asian Languages (ICOSAL) – 4, Dec 3-5, Annamalai University, TamilNadu, 2002.

24. P.Devi Poongulhali, N. Kavitha Noel, R. Preeda Lakshmi, Manavazhahan and T.V. Geetha, "Tamil Text Summarization" International Conference on Knowledge Based Computer Systems, Mumbai, India, December 18-21, 2002.

10. The Team Members

Dr. T. V. Geetha	rctamil@annauniv.edu
Dr. Ranjani Parthasarathi	rp@annauniv.edu
Dr. K. M. Mehata	mehata@annauniv.edu
Dr. Arul Sironmoney	asiro@annauniv.edu
Ms. V. Uma Maheswari	umam_in@yahoo.com
Ms. N. Anuradha	anuradha_n_2000@yahoo.com
Ms. S. Chithrapoovizhi	scpoovizhi@yahoo.com
Ms. J. Deepa devi	deepajayaveer@rediffmail.com
Mr. T. Dhanabalan	tdhanabalan@yahoo.com
Mr. T. Dhinakaran	dhina_t@yahoo.com
Ms. T. Kalaiyarasi	kalaivas@yahoo.com
Mr. A. Manavazhahan	a_manavazhahan@hotmail.com
Mr. G. PalaniRajan	palaniworld@rediffmail.com
Mr. R. Purusothaman	r_purush@hotmail.com
Mr. K. Saravanan	ksaranme@yahoo.com
Ms. M. Vinu krithiga	vinu_krithiga@yahoo.com
Mr. V. Venkateswaran	venkatjee@indiatimes.com
A. Kavitha	avitha_mariyappan@yahoo.com
M. Kavitha	malar_vanam@sify.com

Courtesy: Dr. T.V.Geetha/Ms. Ranjani Parthasarthi
Anna University, School of Computer Science & Engineering, Chennai - 600 025
(RCILTS for Tamil)
Tel: 00-91-44-22351723,22350397
Extn- 3342/3347, 24422620, 2423557
E-Mail: rp@annauniv.edu