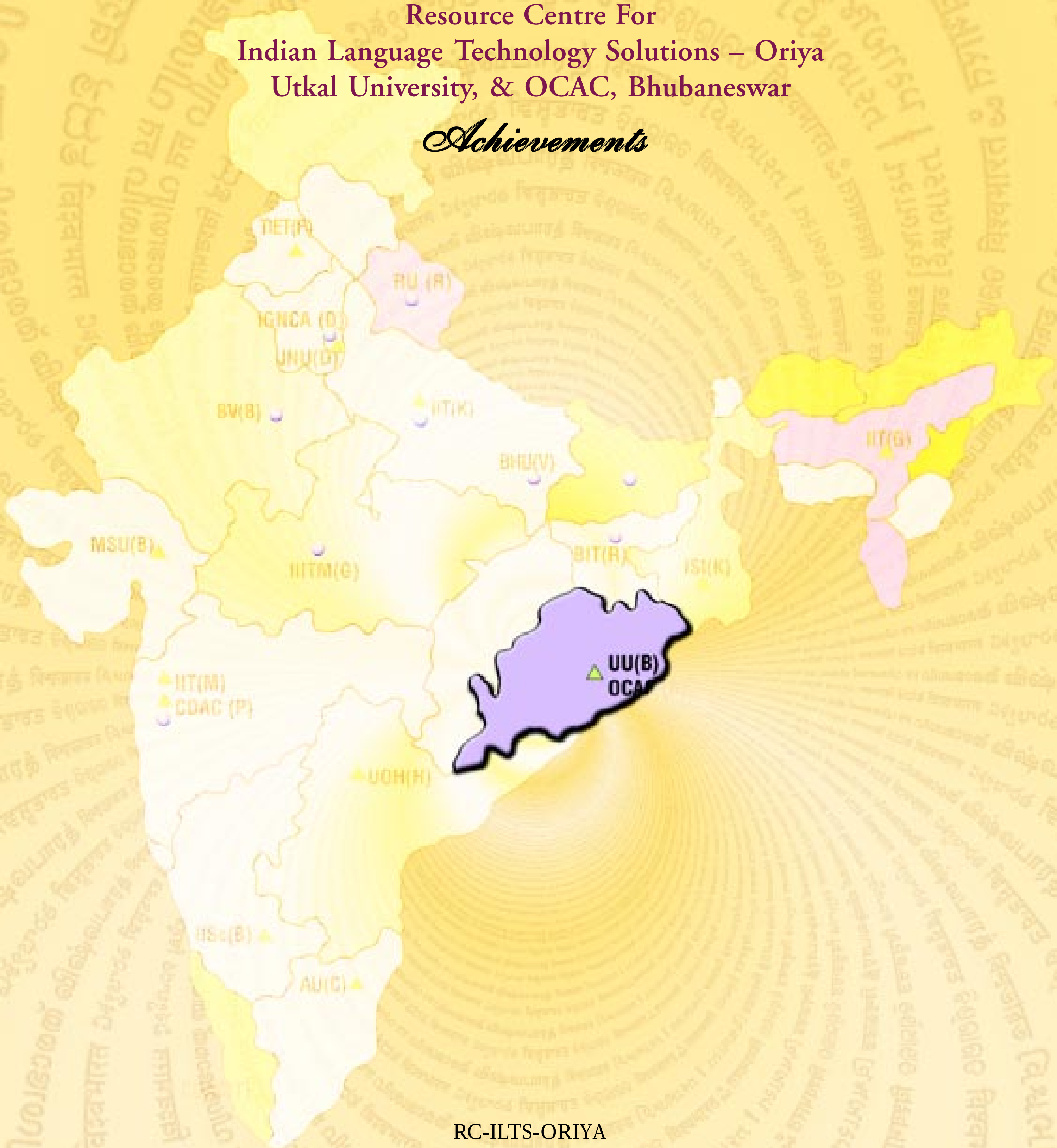
Resource Centre For
Indian Language Technology Solutions – Oriya
Utkal University, & OCAC, Bhubaneswar

Achievements



RC-ILTS-ORIYA
Department of Computer Science & Application
Utkal University, Bhubaneswar, Orissa – 751004
Tel. : 00-91-674-2585518 / 0216 E-mail : sangham1@rediffmail.com
Website : http://www.ilts-utkal.org
&
Orissa Computer Application Centre
Plot No.-N-1/7-D, Acharya Vihar Square
RRL Post Office, Bhubaneswar-751013
Tel. : 00-91-674-2582490/2582850 E-mail : skt_ocac@yahoo.com
Website : http://www.utkal.ernet.in

RCILTS-Oriya
Utkal University, Bhubaneswar

Introduction

Resource Centers (RC) for Indian Language Technology Solutions (ILTS), under Technology Development of Indian Language (TDIL) programs of the Ministry of Communication and Information Technology, Government of India, are established for providing platform to disseminate knowledge to common man through digital unites. One such RC is established in Utkal University, Orissa to handle the issues of Oriya language, the official language of Orissa. For the effective implementation of the idea, several technologies, such as Image Processing, Speech Processing and Natural Language Processing are merged together towards the development of hardware and software under the project for the services of Oriya people. This enable them to be computer literate and thus to be a knowledgeable person. RC-ILTS-Oriya works for the development of tools like

1. Bilingual E-Dictionary English<>Oriya)
2. Oriya Spell Checker
3. Oriya WordNet
4. Oriya Machine Translation System (English – Oriya)
5. Oriya Optical Character Recognition System
6. Oriya Text-To-Speech System

All these softwares are Copy Righted.

Besides this RC team members are also keen in the development of softwares like

1. Oriya Speech-To-Text System
2. Oriya Word Processor with trilingual (Oriya, English, Hindi) Word Processor with spell checking capacity.
3. Sanskrit WordNet.
4. Jagannath Philosophy

1. Intelligent Document Processing (OCR) for Oriya

Document processing is a need for high-level representation of the contents present in the document so that the content can be analyzed and understood clearly and unambiguously. Different approaches are made for optical recognition of characters for different languages like English, Chinese, Japanese and Korean. Very little efforts have been made for the recognition of Indian languages. We have made an attempt for the recognition of alphabetic characters of Oriya language using novel technique, which helps for efficient processing of text (document). Machine intelligence involves several aspects among which optical recognition is a tool, which can be integrated to text recognition and text-to-speech system. To make these aspects effective, character recognition with better accuracy is needed.

The process of Optical Character Recognition of a document image mainly involves six phases:

1. Digitization
2. Pre-processing
3. Segmentation
4. Feature Extraction
5. Classification
6. Post Processing

The digitization phase uses a scanner or a digital camera that divides the whole document into a rectangular matrix of dots taking into consideration the change of light intensity at each dot. The matrix of dots is represented digitally as a two dimensional array of bits. Each dot can be represented by a single bit for b/w image (0-black, 1-white) and for color image, needs 24 bits for its representation. The better the resolution, the better the image.

Preprocessing involves several activities, which transform the scanned image into a form suitable for recognition. These activities are noise clearing, filtering and smoothing, thinning, normalization, and skew correction. But for the recognition of printed characters segmentation plays an important role in the activities of preprocessing. After noise clearing phase, it needs individual characters to be extracted with better approximation so that no characters should loose its important features during the process of extraction. So efficient segmentation algorithms should be employed which will lead to better recognition.

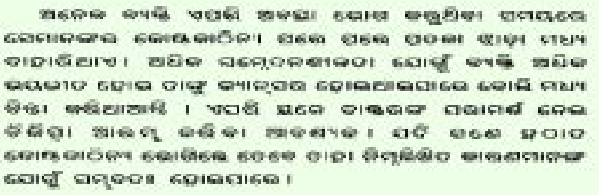
During the segmentation phase, the whole image is analyzed and different logical regions in the image are separated. The logical units consist of texts which has lines thus words thus characters. The error during isolating characters changes their basic shapes. So the characters must be properly extracted so that it will lead to better and accurate representation of the original character. Many algorithms exist for extracting characters from an image. But the problem arises when some characters are connected in the document image. So two characters when connected may be mistaken as a single character, thus erroneous extraction leading to misrecognition.

In the feature extraction phase, the features of individual characters are analyzed and represented in terms of its specialty or uniqueness. These features help in classification and identification of the characters. The feature set needs to be small and the values needs to be coherent for objects in the same class for better classification.

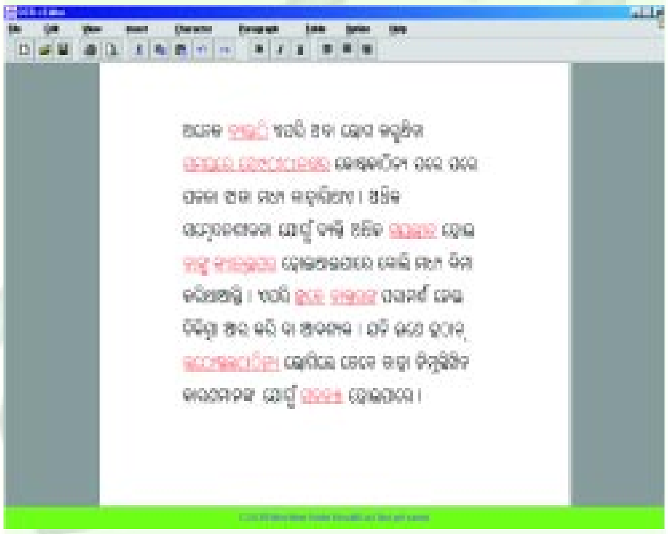
Technology Involved

1. Literature survey done.
2. Gray tone converted to two-tone image by dynamic thresholding on intensity.
3. Skew correction implemented for easy handling of documents.
4. Lines extracted from a document using histogram analysis.
5. Individual characters extracted using region-growing and histogram analysis method.
6. Matras extracted by region analysis.
7. Skeletonization for efficient processing, storing (less memory space allocation) and searching.
8. Connected characters handled by backward and forward chaining of appropriate mask.
9. Features extracted from isolated characters as well as composite characters like jukta.
10. Analysis on Multicolored documents by applying Hue, Saturation and Intensity (HSI) model.
11. The system tested on documents from book (**Bigyandiganta**) and also for some old Oriya books.

The system also integrated to text-to-speech system for Oriya.



(An input image in bmp format)



(The output text in our editor)

2. Natural Language Processing (Oriya)

Natural Language Processing is the technique of engineering our language through machine (the computer) by which we can overcome the language barrier and the difference between man & machine. With that sincere motive we have taken several steps in assisting the computer system behave like a person in exchanging our knowledge. Our efforts are converted to product form and are mentioned below for the interest of researcher of this field and for public interest.

In short our products already developed or under development are Oriya Machine Translation (OMTrans), Oriya Word Processor (OWP) support multilingual, Oriya Morphological Analyser (OMA), ORI-Spell, the Oriya Spell Checker (OSC), Oriya Grammar Checker (OGC), Oriya Semantic Analyser (OSA), ORI-Dic, the bilingual E-

Dictionary (English ó Oriya), OriNet (Word-Net for Oriya), and SanskritNet (Word-Net for Sanskrit) are under developing phase to produce a complete Natural Language Processing System.

2.1 Oriya Machine Translation System (OMTrans)

In OMTrans the source language is English and target language is Oriya. We have developed a parser, which is an essential part of Machine Translation. Our parser is capable in parsing various types of sentence including complex types of sentence such as:

I am going to school everyday for learning.
He said that Ram is a hard working boy.

After the parsing phase the real translation is done. Our translation system is capable in doing the translation of various types of simple as well as complex sentences such as:

I am a good boy.
I am going to school everyday.
I will eat rice.
Who is the prime minister of India?
India is passing through financial crisis.
He told that Ram is a good boy.

Our system is doing the translation of sentences having all types of tense structure. In addition to that our system is also capable of doing the sense disambiguation task based on N-gram model such as:

I am going to <u>bank</u> for money.
I am going to <u>bank </u>everyday.
I am going to <u>bank</u>. I will deposit money there.

In the above examples the meaning of underline word (bank) is decided according to the context of the sentence. Presently it gives very good result for this type of ambiguous sentences.

2.2 Oriya Word Processor (OWP) (Multilingual)

We have developed OWP, which facilitates OSP, OGC and OSA along with multilingual editing. The OWP avails phonetic typing for Oriya character with help facility (Fig-1). All other basic features are also available in the OWP like other word processor (Fig-2). The OWP edits OCR (Oriya Character Recognition) outputs and also the documents of other editors.

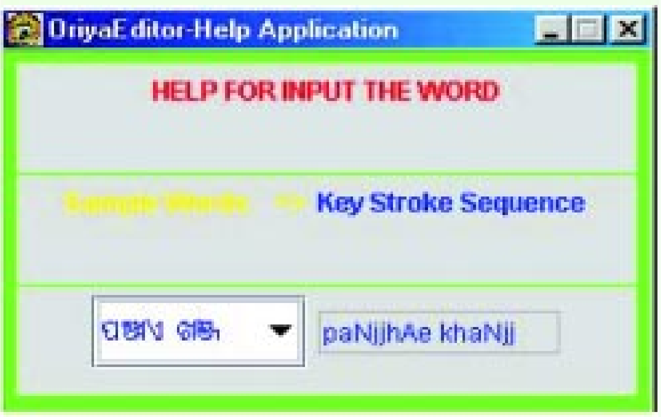

OriyaEditor-Help Application
HELP FOR INPUT THE WORD
Sample Words => Key Stroke Sequence
ପଞ୍ଝାଏ ଖଞ୍ଜି
paNjjhAe khaNjj

Fig 1 : Help option of OWP.

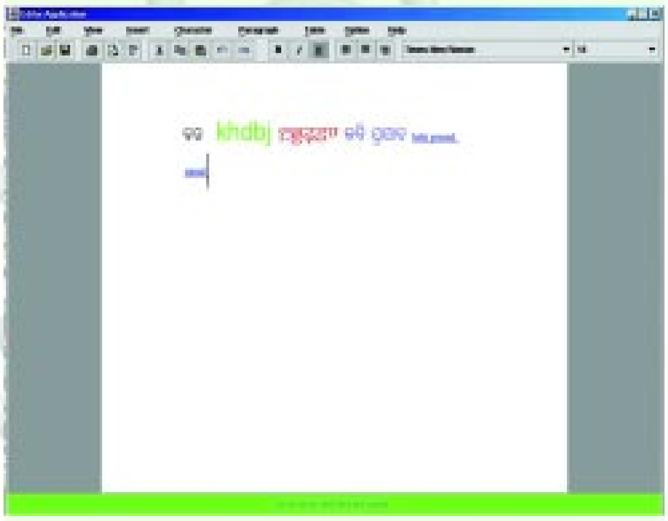
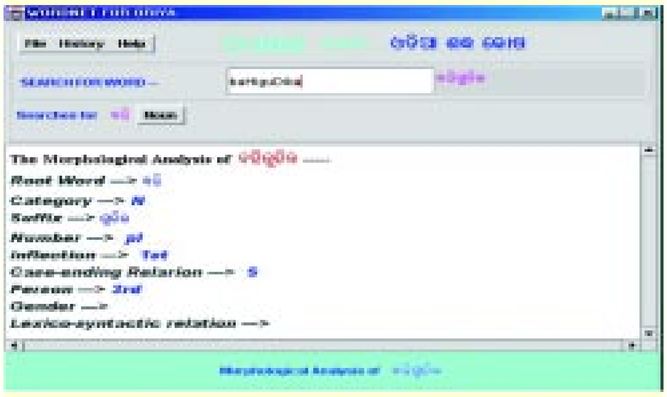


Fig 2 : Multilingual processing of OWP.

2.3 Oriya Morphological Analyser (OMA)

Indian languages are characterised by a very rich system of inflections (VIBHAKTI), derivation and compound formation for which a standard OMA is needed to deal with any type of text in Oriya language. The number of words are being derived from a given root word by some specific syntactic rule. Our OMA deals with morphology of pronoun, number itself (Nominal, Verbal), Number less and prefix (Fig-3). We have developed and implemented some decision tree of each type of morphology, by which our OMA is running successfully. It can help to all the applications involved in MT, OriNet, OSC, and OGC etc.


File History Help

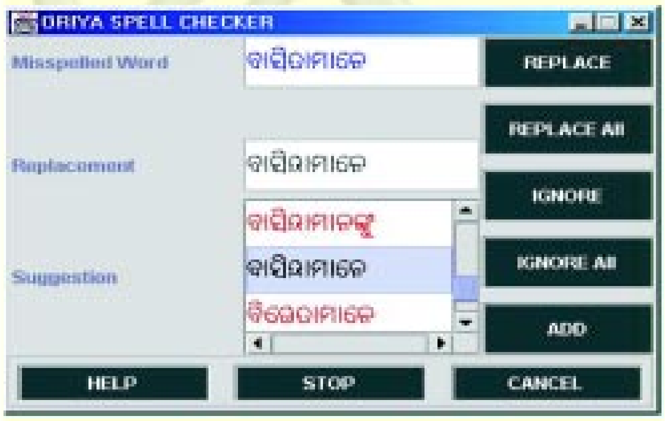
SEARCH FOR WORD -- baHiguDika

The Morphological Analysis of ବହିଗୁଡିକ ----
Root Word --> ବହି
Category --> N
Suffix --> ଗୁଡିକ
Number --> pl
Inflection --> Tat
Case-ending Relation --> 5
Person --> 3rd
Gender -->
Lexico-syntactic relation -->

Fig-3 : Out put of the OMA for the Oriya derived word "baHiguDika" in OriNet.

2.4 Oriya Spell Checker (OSC)

The misspelled words are being taken care of successfully by our OSC. We have developed some algorithms to perform our OSC in order to find out more accurate suggestion for a misspelled word. The searching algorithm of the OSC is so fast that it processes 170000 Oriya words simultaneously for each misspelled word. The words are indexed according to their word length in our word database for effective searching. On the basis of the misspelled word, it (OSC) matches the number of (i) equal character, (ii) forward character and (iii) backward character to give more accurate suggestive words for the misspelled word. Moreover, it also takes help from the Oriya Morphological Analyser for ascertaining the mistakes of derived words. The OSC functions successfully in our Word Processor. This S/W supports both the Windows-98/2000/NT and the Linux as well. Output of the Spell Checker is shown in Fig- 4.


ORIYA SPELL CHECKER

Misspelled Word: ବାସିତାମାନେ

Replacement: ବାସିଥାମାନେ

Suggestion:
ବାସିଥାମାନଙ୍କୁ
ବାସିଥାମାନେ
ବିରେଚାମାନେ

Fig. 4 : Suggestive words for the misspelled word "bAsitAmAne" of OSC.

2.5 Oriya Grammar Checker (OGC)

The S/W for OGC has been developed to determine the grammatical mistakes occurring in a sentence by the help of the OMA and e-Dictionary. It first parses the sentence according to the rule of Oriya grammar and checks the grammatical mistakes. Presently, the OGC functions successfully in our Editor (fig.-5, fig.-6). The OGC -S/W supports both the Windows-98/2000/NT and the Linux as well.

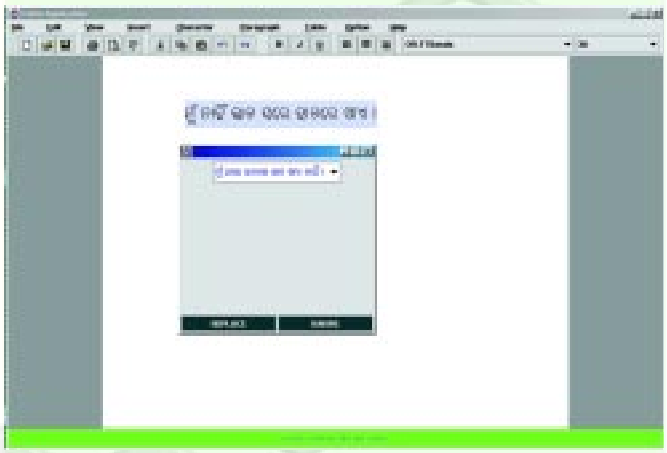


Fig.-5 Suggestion of word sequence of OGC in our OWP.

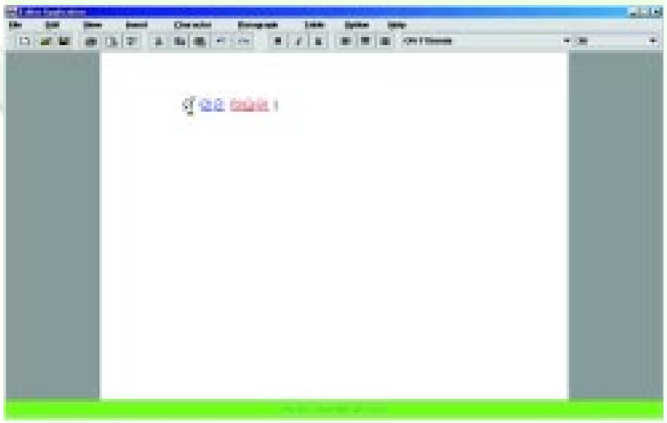


Fig.-6:- Detection of grammar by OGC

2.6 Oriya Semantic Analyser (OSA)

It deals with the understanding procedure of the sentence for which it takes immense help from the KARAKA theory of NAVYA NYAYA philosophy, one of the most advanced epistemological traditions of India. The OSA determines the semantic position of Subject (KARTA), Object (KARMA) etc on the basis of the verb (KRIYA). In other words, it at first links to the verb of the Verb Table (VT) and from verb it links to the subject. Other KARAKAs are determined on the basis of these two linkages (fig.-7). It also takes help from OriNet, OMA and OGC for better understanding. We have worked out 100 verbs in the VT determining their subjective, objective, locative categories etc with respect to these verbs. Presently, with these verbs the OSA functions successfully in our Editor (fig.- 8). The OSA S/W supports both the Windows-98/2000/NT and the Linux as well.

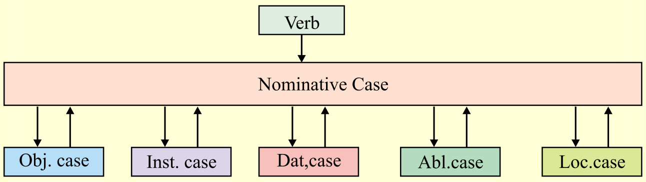


Fig-7:- Semantic Grammatical Model (SGM) and Semantic Extract Model (SEM)

(Here downward arrows represent the SGM and upward arrows along with down arrow from verb represent the SEM where nominative case acts as chief qualifier.)

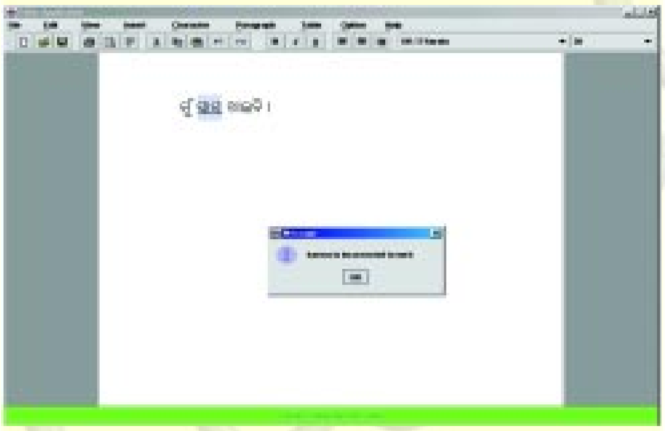
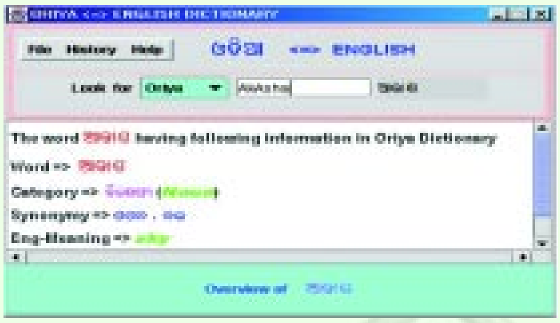


Fig. 8:- Detection of semantic by OSA.

2.7 E-Dictionary (Oriya↔English)

It provides the word, category, synonymy and corresponding English meaning of Oriya as well as English words. The system is successfully functioning over 27000 Oriya words and 10000 English words. Help option also provides the keystroke for each Oriya character in phonetic form. Search engine handles the misspelled word and gives some accurate suggestive words. This S/W supports both the Windows-98/2000/NT and the Linux. We are also in progress to include Hindi and Sanskrit words in this system for the benefit of the users.



ORIYA <-> ENGLISH DICTIONARY

File History Help

ଓଡ଼ିଆ <=> ENGLISH

Look for Oriya AkAsha [illegible]

The word [illegible] having following information in Oriya Dictionary

Word => [illegible]

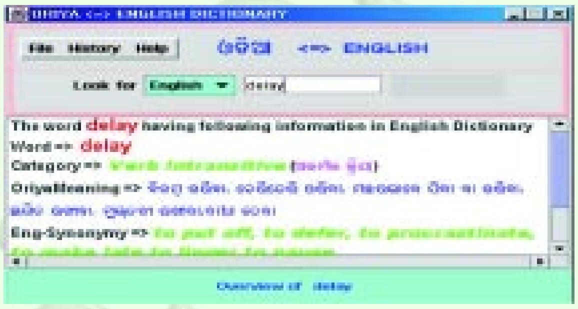
Category => [illegible]

Synonymy => [illegible]

Eng-Meaning => [illegible]

Overview of [illegible]

Fig. 9: Overview of the Oriya Word "AkAsha."



ORIYA <-> ENGLISH DICTIONARY

File History Help

ଓଡ଼ିଆ <=> ENGLISH

Look for English delay

The word delay having following information in English Dictionary

Word => delay

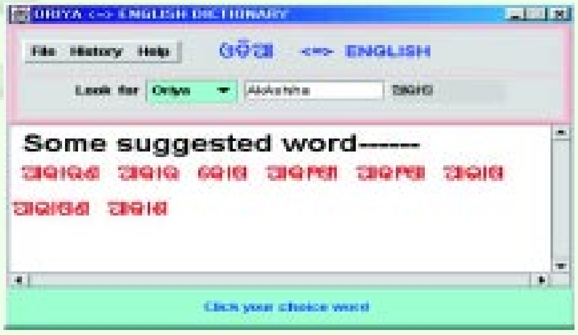
Category => Verb Intransitive ([illegible])

OriyaMeaning => [illegible]

Eng-Synonymy => to put off, to defer, to procrastinate, [illegible]

Overview of delay

Fig. 10 : Overview of the English Word "delay".



ORIYA <-> ENGLISH DICTIONARY

File History Help

ଓଡ଼ିଆ <=> ENGLISH

Look for Oriya AkAshA [illegible]

Some suggested word------

[illegible]

Click your choice word

Fig. 11 : Search Engine (SE) handles misspelled word and gives suggestion.

2.8 OriNet (WordNet for Oriya)-Online lexical dictionary/ thesaurus.

One of the major problems in the implementation of Natural Language Processing (NLP) or Machine Translation (MT) is to develop a complete lexical database containing all types of information of words. There are some difficulties in deciding that what information should be stored in a lexicon and even greater difficulties in acquiring this information in proper form. The OriNet system is designed on the basis of multiple lexical database and tools under one consistent functional interface in order to facilitate systems requiring syntactic, semantic and lexical information of Oriya language. The system is divided into two independent modules. One module is developed to write the source files containing the basic lexical data and these files are taken as the input for OriNet system. Lexicographer takes care the major work of this module. Second module is a set of programs by which it accepts the source files, processes it to display for the user and also provides different interface to use other applications. System has been designed using Object-Oriented paradigm according to Oriya language structure with over 1100 lexical entries, which allow flexibility, reusability and extensibility. It also provides X-windows interface to access the data from the OriNet database as per user's requirement. It can be widely used in different application like, (i) Word Sense Disambiguation (WSD) in Oriya Machine Translation (ii) Oriya Grammar Checker (OGC) and (iii) Oriya Semantic Checker (OSC). Moreover, it helps as a lexical resource for Oriya learners and also for expert scholars who are involved in the research in NLP. The system also consists of Oriya Morphological Analyzer (OMA), which takes care of any type of word such as root word or derived word and also provides syntactic information of the word. This S/W supports both the Windows-98/2000/NT and Linux. Presently, we are adding more and more lexical entries in the source file and developing different applicable programs for use in wider range.

Architecture and Output

The architecture of the OriNet system is divided into five parts (Fig.-12) such as Lexical Source File (LSF), OriNet Engine (OE), OriNet Database (OD), Morphological Analyser (MA) and User Application (UA). The LSF is the collection of different sorted files according to their syntactic categories, which are taken as the inputs to the OriNet system. The OE is a set of programs, which compiles the lexical source files into a database format that facilitates machine-retrieved information in proper manner. It also acts as the CPU of the OriNet system. It is used as a verification tool to ensure the syntactic integrity of lexical files. All of the lexical files are processed together to build the final OD. It also provides sufficient information as per user's query or other application with the help of MA.

At the heart of the OriNet system is the OD, which stores the data in simple ASCII format, several easy-to-use interfaces have been devised to cope with varied user requirements and the raw data extracted from the database are well formatted for display purpose. MA takes care syntactic analysis of any types of lexicon and also provides the root words with other grammatical information. X- windows interfaces are also designed to access the database with different types of options (Fig.-13).

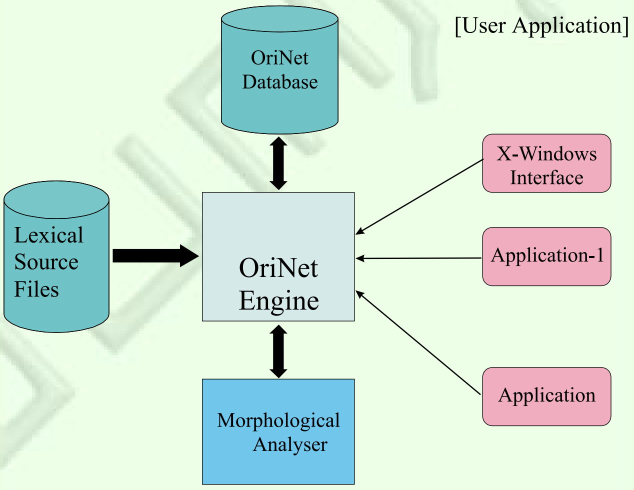
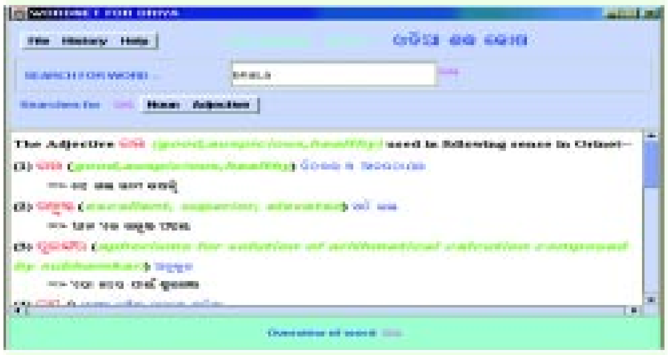


Fig. 12:- Architecture of the OriNet System



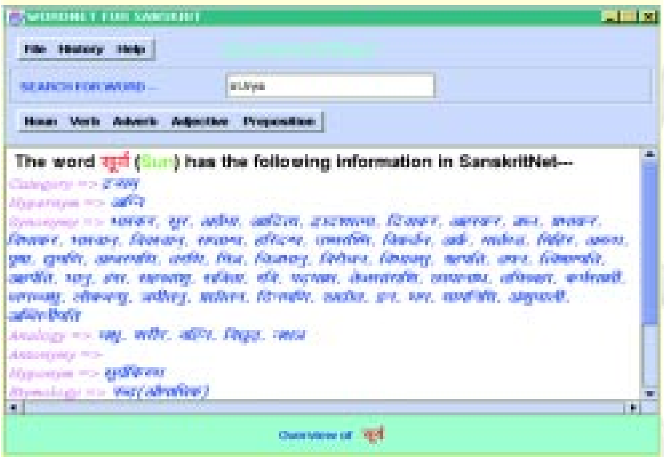
The Adjective [illegible] used in following sense in Orinet—

Fig. 13:- Overview of the word bþmþ (good).

2.9 SanskritNet (Word-Net in Sanskrit)-Online lexical dictionary/thesaurus.

Sanskrit language is the base for most of the Indian languages. It has link with foreign languages too. To study this language and also to use it for knowledge enhancement, effective Machine Translation (MT) from one language to the other is necessary for which online lexical resources are needed. Towards making such resources we have tried to develop a Word-Net for Sanskrit language using the *Navya-NyAya* (नव्यन्याय) Philosophy and Paninian grammar. Besides Synonymy, Antonymy, Hypernymy, Hyponymy, Holonymy and Meronymy we have also introduced Etymology and Analogy separately as they play important roles in *Navya-NyAya* (नव्यन्याय) Philosophy, which is the specialty on this Sanskrit Word-Net for a better classification of words.

Presently, we have defined and analysed 300 Sanskrit words (200 nominal words and 100 verbal words) in the SanskritNet. On the basis of the analysis and the definition of these words, we are on progress to design and implement the SanskritNet and other allied applications as well. Here the prototype model of the SanskritNet is displayed (Fig.-14).



The word सूर्य (Sun) has the following information in SanskritNet---
Category => [illegible]
Hypernym => [illegible]
Synonymy => [illegible]
Analogy => [illegible]
Antonymy =>
Hyponym => [illegible]
Etymology => [illegible]

Overview of सूर्य

Fig. 14: - Output of the Nominal word "सूर्य" (Sun) in SanskritNet

3. Speech Processing System For Indian Languages

In the age of fast technology, where information travels at the speed of light, we still rely on feeding our text inputs in typographical manner. Speech Processing System is an approach to provide a speech interface between user and the computer. Basically our system zeros in on *Hindi*, *Oriya* and *Nepali* languages at the Resource Centre.

Broadly the system is classified into two sections.

- Text To Speech (TTS)
- Speech To Text (STT)

Aim

To design a system/algorithm, which works efficiently, naturally and utilizes memory as less as possible.

Progress

We have developed TTS for *Oriya* and *Hindi* language and the TTS for Nepali language is in progress.

3.1 Text to Speech (TTS)

As the name signifies this system provides an interface through which a user enters certain text/document and it is the software that reads it as natural as a human. The basic approach followed here is, first to analyse the document (language, font etc.), and then extract words from the text, try to parse individual words into vowels and consonants respectively. Then corresponding to these vowels and consonants existing (previously stored in the database) ".wav" files are concatenated and played.

Technologies Behind :

- Creating the wave file database.

For creation of such a database we studied a lot of recorded words and sentences and try to break them into vowels and consonants by minute hearing. Then we analyse those cut pieces and store the appropriate and generalised form in the database.

- Parse extracted exact words of a given sentence?

The same words in different sentences have different stress due to its position in the sentence. Appropriate hidden vowels are detected form the words extracted. For example considering a word "ସମୟ" (*SAMAYA*) in *Oriya* is parsed as follows :-

ସ୍ + ଅ + ମ୍ + ଅ + ୟ୍ + ଅ

The format of vowel and consonant break point is shown in the fig1.

Again considering the same word in Hindi i.e., *"समय"* (*SAMAY*) is parsed as follows :-

स् + अ + म् + अ + य् + अ

- Choosing of appropriate '.wav' file from the database.

Considering the above example, may be the vowels we get after parsing are the same as 'ଅ' (for *Oriya*) or 'अ' (for *Hindi*), but it is not exactly the same 'a.wav' we concat in every case. Thus, we analyse vowels broadly in three categories as ma:tra: in

- Beginning
  Middle
- End

And this is observed that the duration of ma:tra:s say 'ଅ' here, varies from each other, i.e. 'ଅ' in middle is not the same as that in the end. Again accordingly we need to get the appropriate ".wav" files from the database.

As observed in the example

i) The durations of 'ଅ'(*Oriya*) are as follows:

Starting ma:tra:	- 0.065 sec
Middle ma:tra:	- 0.105 sec
End ma:tra:	- 0.116 sec

ii) The durations of 'अ'(*Hindi*) are as follows :

Starting ma:tra:	- 0.063 sec
Middle ma:tra:	- 0.078 sec
End ma:tra:	- 0.020 sec

- It is observed that concatenation of the wave files is not that natural as expected. This is due to the certain transitions between the characters in the actual pronunciation. Thus we are developing a robust algorithm for the generation of naturalness in the TTS output.

Applications

# Helpful for the blinds and illiterates.

# Airways and Railways announcement system.

# Initial phase of Speech-to-Speech conversion system.

3.2 Speech to Text (STT)

The speech processing is a pattern recognition problem. The recognition of speech is defined as an activity whereby a speech sample is attributed to a person on the basis of its phonetic-acoustic or perceptual properties. In our approach we study the nature of spoken words by different speakers. From a continuous sentence the word boundaries are detected and the nature of utterance of individual consonants and vowels are marked to study their behaviour for a particular speaker. We use word spotting techniques and the variations of pitch and intonation are marked. The algorithm makes the use of the properties of F_0 contours such as declination tendency, resetting and fall- rise patterns in the utterance. Once the patterns are obtained it is mapped with a particular character and the proper text output is obtained. The boundary and parameter detection is shown in fig2. After obtaining the parameters of an uttered word, they are standardise to maintain the speech database. We also provide some tolerance in each parameter, so that little variation in the utterance doesn't affect the recognition of the word for text conversion. Form each word the characters are obtained and when a speaker utters, the corresponding character is mapped to the text and the output is obtained.

wah kitna sunder hai !

वह कितना सुन्दर है !

The stress on the vowels (rising and falling) are shown in the pitch curve. The spectrograms of the uttered words show the content frequency and intensity of utterance. The word boundary is detected by formant analysis. Though F_0 plays a major role. The cepstral component of the continuous speech puts major emphasis in the utterance. Hance taking the word boundary into consideration the speech database show around 73% accuracy for hindi and 90% in Oriya speech recognition.

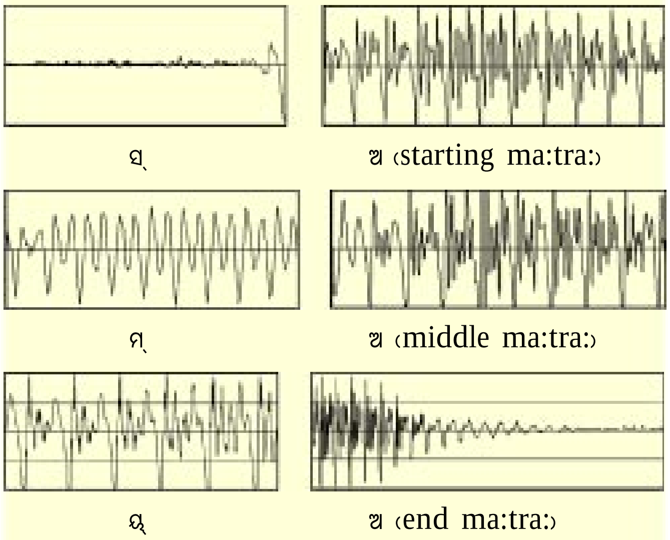
Application

# Speech password for security in banking sector.

# Foreign sic experts need for speech recognition & identification of criminals.

# For the blind and illiterate people, provides an intuitive interface with machine.

# Final phase of Speech-to-Speech conversion System.



ଘ୍

ଅ (starting ma:tra:)

ମ୍

ଅ (middle ma:tra:)

ଝ୍

ଅ (end ma:tra:)

Fig. 1

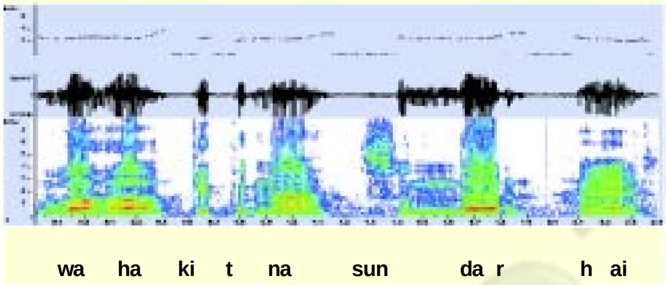


Fig. 2

4. The Team Members

Khyamanidhi Sahoo	khyamanidhi@yahoo.com
Hemanta Kumar Behera	kumar_hemanta@yahoo.com
Pravat Chandra Satapathy	satapathy_pravat@rediffmail.com
Suman Bhattacharya	suman_bh1@yahoo.com
Ajaya Kumar senapati	ajayasenapati@rediffmail.com
Pravat Kumar Santi	pksanti@rediffmail.com
Gour Prasad Rout	rout_g_prasad@yahoo.com
Kabi Prasad Samal	kabi_prasad@rediffmail.com
Krushna Pada Das Adhikary	krushnapada@rediffmail.com

Courtesy : Prof. (Ms) Sanghmitra Mohanty
Utkal University
Department of Computer Science & Application
Vani Vihar, Bhubaneshwar – 751 004
(RCILTS for Oriya)
Tel: 00-91-674-2585518, 254086
E-mail: sangham1@rediffmail.com)

RCILTS-Oriya
OCAC, Bhubaneswar

1. Product Developed at OCAC and the Target date of its Commercialization

1.1 Oriya Spell Checker

Oriya Spellchecker has been developed in Linux platform. It consists of 60,000 base words. These base words can be manipulated and stored in dictionary in a scientific manner. Root words can also be manipulated. This Spellchecker is incorporated in our Oriya word processor Sulekha. This project is developed for the smooth use of novice users.

Here we are concentrating with our suggestion and checking of words. The checker checks the pattern string and suggests to the user more accurate and required suggestion. Like English Indian languages are different in their nature. Indian languages are complex in their nature. For that we have designed a checker and that checks the pattern string. Words are stored in side the dictionary file line by line. Each line is terminated through an escape sequence of new line character. All these Oriya words are stored inside a file having extension .txt.

Features

Spellchecker consists of several facilities like Prefix, Suffix, Add, Ignore, Change, Spellcheck, Suggestion and Cancel. Suggestion is accurate and match more closest word to the user requirement. It checks more words within a fraction of seconds. Searching time is minimum and complexity is less. Spellchecker can be incorporated in any application for example it can be incorporated in a word processor or an editor or a browser. It uses ISCII data stored in a file. ISFOC for display. Gets the input ISFOC data displayed the ISFOC data.

Commercialization

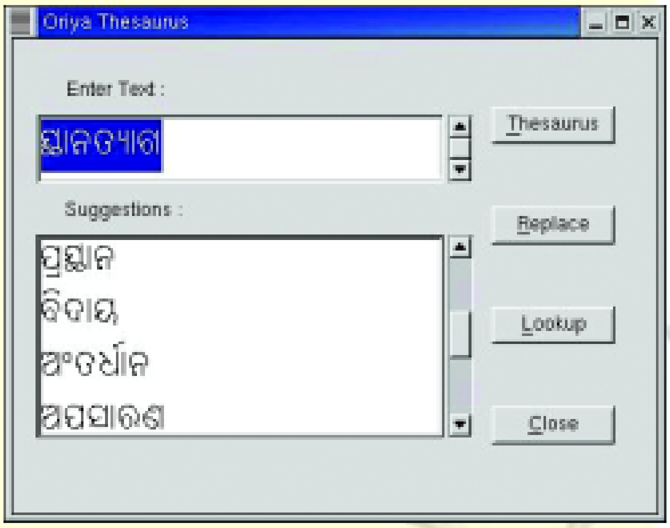
Windows version has been completed and commercialized and incorporated in LEAP Office of CDAC Linux version Completed

1.2 Thesaurus in Oriya

Oriya Thesaurus has been developed in Linux platform. It consists of 40,000 base words. These base words can be manipulated and stored in dictionary in a scientific manner. Root words can also be manipulated. This Thesaurus is incorporated in our Oriya word processor Sulekha. This project is developed for the smooth use of novice users. It is a tool for correct documentation. Words are stored in side the dictionary file line by line. Each line is terminated through an escape sequence of new line character. All these Oriya words are stored inside a file having extension .txt.

Features

Thesaurus consists of several facilities like hesaurus, Replace. Suggestion is accurate and matches closest word to the user requirement. It checks more words within a fraction of seconds. Searching time is minimum and complexity is less.Thesaurus can be incorporated in any application for example it can be incorporated in a word processor or an editor or a browser.



ସ୍ଥାନଚ୍ୟୁତ

ପ୍ରସ୍ଥାନ
ବିଦାୟ
ଅଂତର୍ଧାନ
ଅପସାରଣ

Commercialization
Completed

1.3 Bilingual Electronic Lexicon

The dictionary for administration and official correspondence has been entered for Oriya-English and English-Oriya. It is proposed to bring out an Electronic Dictionary as a commercial product which will have features such as phonetic key board and user friendly functions for "add", "delete", "view" etc. After the availability of text to speech technology in Oriya the same will be integrated to this product for providing help in correct pronunciation of words. Lexical inputs from other domains such as culture, business, science are being collected.

Commercialization
Words have been entered. As a product with more words will be completed in June 2003.

1.4 Corpus in Oriya

A large corpus has been developed consisting of nearly 8 million words for analysis and use in spell checker and thesaurus. Corpus creation has been taken as a continuing activity as a part of this project.

Commercialization
It will be used for research and use in other products and is not for commercialization,. (completed)

1.5 Bilingual Chat Server

Oriya Chat is a Standalone Chat Application in Oriya Language developed in JAVA. After connecting to the Server, one can chat in one's favorite room, send direct messages with emotion icons to other online users as well as send public messages.

Features
User Log in to Chat Server with unique ID, Connect to the Server, with or without firewall/proxy, Phonetic Keyboard Layout to write in Oriya,Send Oriya text with Emoticons (images), Private and Public Chat With Friends, Domain specific Predefined Public Rooms.

System Description
Oriya Chat has a client and a server module. Serversocket and SocksSockets are used for communicating with remote machine. SocksSocket is a CustomSocket that extends Socket to be used with the firewall or proxy.

Commercialization
Completed

1.6 Net Education

A Net Education System (NES) is being developed using Bilingual Chat Server and Mail Server. The educational Content is being created for providing students, public for education on various aspects through the net.

Commercialization

The Chat Sever with Interface has been completed. A full fledged system with content will be completed soon.

1.7 XML Document Creation and Manipulation

XML document creation technology in English and Oriya language has been developed. It incorporates users defined tags, structures the document like a database providing real world meaning of the content. This provides more flexibility to the presentation of the same document in different style sheets and same style sheet for different documents. Also queries can be processed on the contents of the XML document. Content on Tourism has been created in XML. Agricultural content useful for farmers are being created.

Commercialization

A few documents have been completed. More work is in progress to be completed by August 2003.

1.8 Oriya OCR

OCAC is collaborating with ISI Calcutta to develop and commercialization of Oriya OCR . All the language inputs has been provided to ISI Calcutta. The software developed by ISI Calcutta is being tested by OCAC. OCAC will commercialize the Oriya OCR with know how from ISI Calcutta.

Commercialization

Work in collaboration with ISI, Calcutta is in progress. Will be completed in June 2003

1.9 Oriya E-Mail

A majority of Indian population cannot speak, read and write in English language. Oriya language based tools helps the common man to benefit from e-governance and other applications of Information Technology.

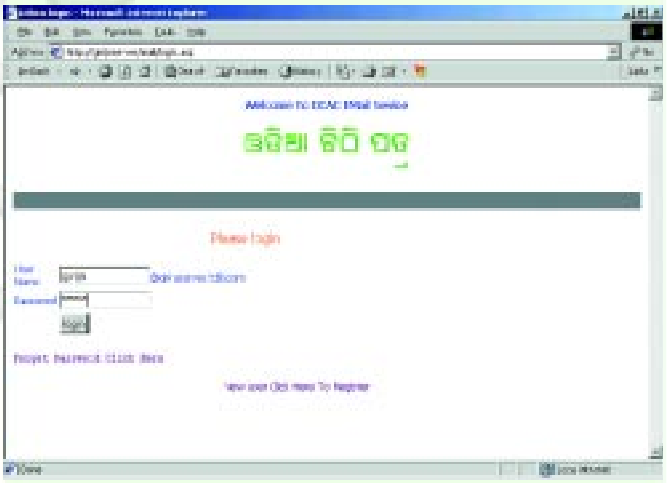
In order to bridge the digital divide it is necessary to provide Internet application such as e-mail and chat in Oriya language for use by the common man. Varta is an e-mail solution, which enables people to send and receive mail in Oriya language. It uses dynamic version of the OR-TTSarala font, created using Microsoft WEFT-III. Use of dynamic font allows the client to read and compose the mail without downloading the Oriya font. The system has been developed using ASP technology, Java, JavaScript and HTML. Varta uses default SMTP Server on the Internet Information Server 5.0 of Window 2000 Server.

Features

Most of the basic e-mail features are available in Varta 0.2, which includes : Sending, Replying, Forwarding and Reading Oriya Mail using dynamic Font, Online Registration of unique user mail account, Password Security with Online Help to retrieve forgotten Password, Phonetic keyboard typing in Oriya, Online typing help through a keyboard map to dynamically select appropriate keys.

Phonetic Keyboard Engine

The Oriya keyboard has been designed as per Indian phonetic standard. It is implemented as an applet in Java language. It is compiled in Java 1.0 JDK so that it runs in MS Internet Explorer 5.0 onwards.



Welcome to OCAC EMail Service

Please login

User Name

Password

login

Forgot Password Click Here

New user Click Here To Register

Commercialization

Completed

1.10 Oriya Word Processor with Spell Checker under LINUX

For the first time an Oriya word processor has been developed on Linux operating system. It runs on any X-windows system. Two Keyboard engines (Inscript keyboard engine and Phonetic Keyboard

Engine) have been incorporated for the ease of typing. All the texts will be displayed in GLYPHCODE. For display of a stored ISCII file, a converter engine automatically generates GLYPHCODE equivalent of the file.

Features

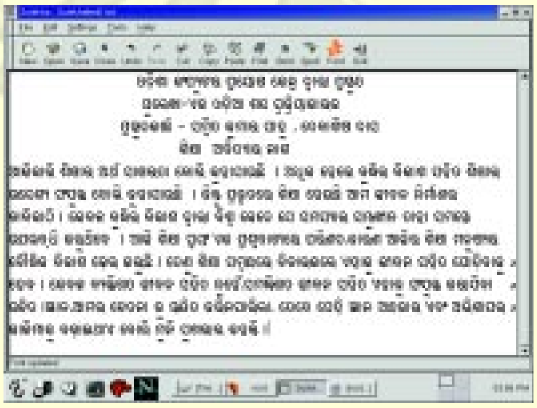
Most of the basic editing features available in standard word processor have been incorporated which includes : Opening new file, saving a file, closing window, cut, copy, paste, Single level Undo and Redo is provided, Keyboard engine for typing in Inscript layout, Keyboard engine for typing in Phonetic layout, The ISCII to GLYPHCODE conversion, The GLYPHCODE to ISCII conversion, Storing content of a file in plain text as well as in ISCII format, Other features such as Find, Find and Replace, Go To Line, File Information, A Spell Checker incorporated within the word processor.

Keyboard Engine

Two Keyboard engines have been designed. One according to Indian script (Inscript) standard. Another according to Phonetic standard. All the consonants, vowels and conjuncts and special symbols etc are displayed in Glyph Code format. Invalid word combinations according to Oriya Grammar are disallowed. Pressing the Scroll Lock Key activates the Keyboard Engine.

Converter Engine

An Export file option is provided to convert 8-bit glyph code to 8-bit ISCII code format. The 8-bit ISCII code can also be converted back to 8-bit glyph code.



Commercialization

Completed

1.11 Computer Based Training (in Oriya)

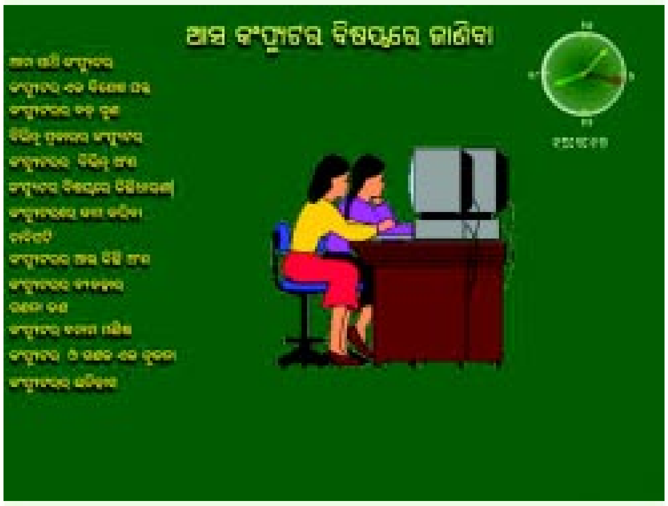
It is well known that multimedia CDs are highly effective medium of learning in many disciplines. The integration of text with hypertext links, speech, graphics, animation and video in an interactive mode makes the lessons easy to visualize and understand for self-learning. Computer Based Tutorial CDs for school children have been developed to support the classroom teaching with interactive self-learning program.

Features

Informative content with animated illustrations, Interactive quiz for evaluation of subject understanding.

Description

This project is aimed at the development of tutorial CDs based on School Curriculum, in Oriya language. The content is based on textbooks prescribed by Board of Secondary Education, Govt. of Orissa. The lessons are illustrated with text, speech, and animation with facility to navigate through various chapters using hyperlinks. At the end of a chapter, the on line quiz helps in self-evaluation by the student. It is based on Windows 98 Operating System and Tools used are Macromedia Authorware 6, Macromedia Flash 5.0, Adobe PageMaker 6.5, Adobe Illustrator 9.0. The minimum System Requirements for Installation is Pentium II or higher, 128 MB or more RAM, Internet Explorer 5.5 or more



Commercialization

Two CD have been **completed** .Write up with colour brochure enclosed. Work in more subject is in progress.

1.12 Oriya Language Based E-Governance Applications

A number of software modules have been developed for enabling the common citizens to access government information and services in Oriya and English languages from the Kiosks. A citizen can submit grievance petition; can apply for one or more certificates (such as birth, death, caste, income etc; down load the application forms and know the rates of consumer commodities etc. These modules have been used as a pilot project on improving citizens' access to information in Kalahandi district of Orissa.

Commercialization

Development of a number of applications has been **completed** and integrated with the project on Citizens' Access to Information.

2. TOT done for various projects so far

1. Spell Checking Grammar rules in Oriya Language has been Transferred to CDAC.

2. Documents on Oriya Language Standards and Principles have been prepared and for UNICODE and sent to MIT.

3. The draft for the preparation of National Language Design Guide for Oriya has been prepared and submitted to MIT

4. Oriya Script with Grapheme details have been provided to ISI, Calcutta for Oriya OCR.

3. Training programmes run for the officials for the State Govt.

OCAC had earlier trained 70 persons in various government offices on Oriya Language Word Processing using LEAP Office in which the Oriya Spell Checker has been incorporated. The last formal training course on " Word processing in Oriya Language using Leap Office and ISM " was conducted during January 17 - February 12 2002. in which only four officials attended the course. Again a course has been organized for 106 staff of the Board of Revenue during 29-07-02 to 3-08-02 (first batch) and 5-08-02 to 10-08-02 (second batch). So far 180 officers have been trained. We are developing course material on a CD for training on Oriya Language.

4. Core Activities

4.1 Web hosting of Oriya Language Classics

Three well known Oriya classics :

(1) "Chha Mana Atha Guntha" by Fakir Mohan Senapati,

(2) "Chilika" by Radhanath Ray and

(3) "Tapaswini" by Gangadhar Meher - **have been hosted** on Internet at http://www.tdil.utkal.ernet.in

4.2 Hosting of Web Sites of Govt. Colleges in Orissa

Web sites **have been hosted** for the Khallikote College, Berhampur www.khallikotecollege.utkal.ernet.in and Ravenshaw College, Cuttack. www.ravenshawcollege.utkal.ernet.in The pages include information on the college as well as forms for students feed back. More dynamic information will be supported by the web sites in future, which will be undertaken for many colleges in Orissa.

5. Products proposed to be developed

1. CBT for all classes (std – 4th to 10th)
2. OCR
3. Upgradation of spell checker
4. Grammar module development
5. Machine Translation System (Oriya –English)
6. Net Education System in Oriya Language

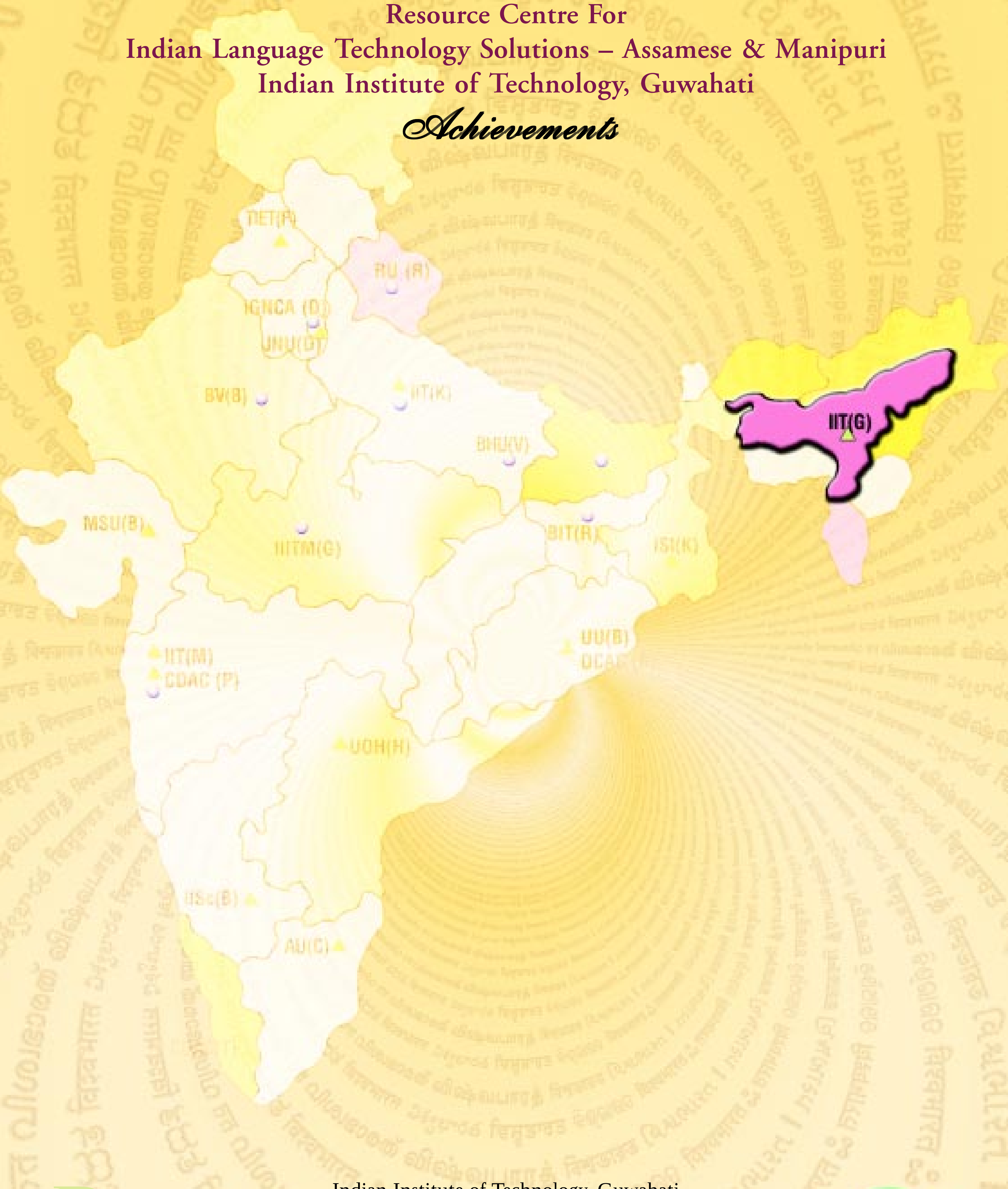
6. The Team Members

Biswa Ranjan Sahoo	biswa_litu@yahoo.com
Sambit Kumar Sahu	sam_sahu@yahoo.com
Girija Shankar Sarangi	_sarangi@rediffmail.com
Sunil Kumar Panda	k2_sunil@yahoo.com
Sarita Das	sarita_das@yahoo.com
Subhendu Kumar Mohanty	ohanty_subbu@yahoo.com
Sanjay Kumar Dey	sanjay_k_dey@yahoo.com

Courtesy : Shri S.K. Tripathi
Orissa Computer Application Centre
OCAC Building, Plot No. 1/7-D,
Acharya Vihar Square, RP-O,
Bhubaneswar – 751 013
(RCILTS for Oriya)
Tel: 00-91-674-2582484, 2582490,
2585851, 2554230 (R)
E-mail: saroj@ocac.ernet.in

Resource Centre For
Indian Language Technology Solutions – Assamese & Manipuri
Indian Institute of Technology, Guwahati

Achievements



Indian Institute of Technology, Guwahati
Department of Computer Science & Engineering, Panbazar
North Guwahati, Assam-781031 India
Tel. : 00-91-361-2691086 E-mail : tdil@iitg.ernet.in
Website : http://www.iitg.ernet.in/rcilts