5.1 Bangla Spell-Checker

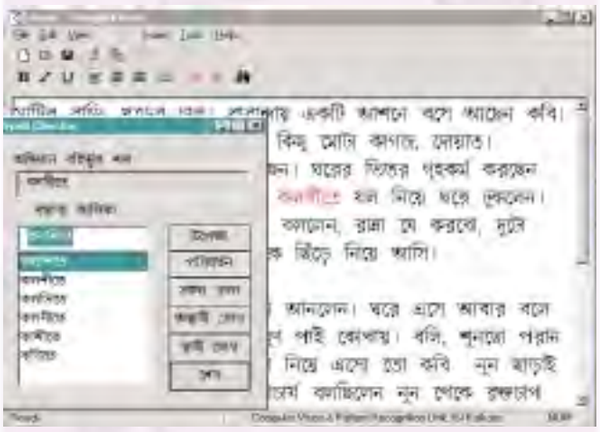
5.1.1

- **Name of the Technology :** ***Bangla Spell Checker (Banan Parikshak)***
- **Nature of Technology :** Knowledge tool (Automatic Spell Checking and correcting Bangla words in Bangla editor)
- **Level:** (Product / Technology / Sub-system) Product Software ready for transfer for commercialization
- **Technical Description of the Technology / Product** including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product : Bangla spell checker can act in both offline and online modes. It has a Graphics User Interface via an editor (described below). Whenever the user types Bangla text, it checks for wrong spelling and gives suitable suggestions. On the other hand, one can run this software on previously typed material such as an OCR document. All misspelled words are captured by the system. For a single-error word, intended word is found within top four words in the suggestion list. Words having more than one error are so captured and for most of them intended words are in the upper half of the suggestion list. However, suggestions cannot be given on some inflected words. It has facility to add new words in the dictionary against which spellings are checked.

  The spell-checker is a tool to detect errors in Bangla words and correcting them by providing a set of correct alternatives which includes the intended word. An erroneous word can belong to one of two distinct categories, namely, nonword error and real-word error. Let a string of characters separated by spaces or punctuation marks be called a candidate string. A candidatestring is a valid word if it carries a meaning. A meaningless string is a non-word. By real word error we mean a valid but not the intended word in the sentence, thus making the sentence syntactically or semantically ill-formed or incorrect. In both cases, the problem is to detect the erroneous word and either suggest correct alternatives or automatically replace it by the appropriate word. In this spell-checker, we consider only the non-word error.

  Supporting platform(s): Windows 95/98/2000/ NT

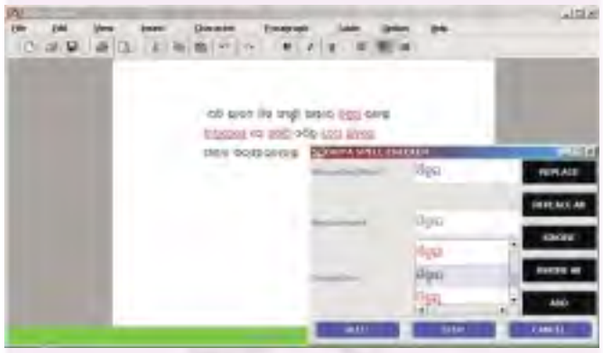
  Memory requirements: 2MB or above
- **Representative Snapshot / screenshot of the Technology / Product:** Screenshot of Application of Spell Checker in Bangla Editor



- **Scalability / Portability / Expandability:** Portability: The system can run in WIN9X platform.
- **Readiness of Transfer of Technology (ToT) :** Yes
- **Availability of documentation:** Yes
- **Testing of the Product / Technology :** Locally tested
- **IPR / Open-source :** Not open source
- **Potential Beneficiaries:** Typists, Linguists, Students, Teachers, Educationists, Office Personnel, Corpus Developers etc.
- **User–agency tie-up :** None
- **Name and address of the Resource Person:**
  B. B. Chaudhuri, Professor & Head,
  Computer Vision & Pattern Recognition Unit
  Indian Statistical Institute
  203 Barrackpore Trunk Road
  Calcutta 700 108, INDIA
  E-mail : bbc@isical.ac.in
  Web : http://www.isical.ac.in/rrc-bangla

5.1.2

- **Name of the Technology : *ORI-SPELLCHECK (Oriya Spell Checker)***
- **Nature of the Technology:** Knowledge tool (Application oriented)
- **Level:** (Product / Technology / Sub-system): Product
- **Technical Description of the Technology / Product** including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product: Error detection and automatic or manual correction for misspelled words is taken care of successfully by Oriya Spell Checker (OSC). We have developed some algorithms to perform our OSC in order to find out more accurate suggestions for a misspelled word. The searching algorithm of the OSC is so fast that it processes nearly 170000 Oriya words for each misspelled word. The words are indexed according to their word length in our word database in order for effective searching. On the basis of the misspelled word, it (OSC) takes into account the number of (i) matching characters, (ii) corresponding matching characters in forward direction and (iii) corresponding matching characters in backward direction to give more accurate suggestive words for the misspelled word. Moreover, it also takes help of the Oriya Morphological Analyser (OMA) for ascertaining the mistakes of inflection (VIBHAKTI), derived and compounding words. Our Correction engine corrects automatically the error like phonetically similar characters; OCR's miss recognised characters with the help of confusion matrix and misspelled word correction by using most phonetically confused word database through n-gram model. The OSC is running successfully in our Word Processor. Using this technique, we have also developed **Hindi Spellchecker** (270000 words) and **English Spellchecker** (20000 words) for our Word Processor (WP). The word files are stored in ASCII format and supports OR-TTSarala (Oriya), DV-TTYogesh (Hindi) and any font in English. This S/W is designed using Java and Java Swing (Front-end) for both the Windows-98/2000/NT and the Linux O/S.
- **Representative Snapshot / screenshot of the Technology / Product:**



- **Scalability / Portability / Expandability:** All
- **Readiness of Transfer of Technology (ToT):** Yes
- **Availability of documentation:** Yes
- **Testing of the Product / Technology:** Under progress
- **IPR / Open-source:** IPR (No- SW-1176-2003)
- **Potential beneficiaries:** Common man & Research Scholar
- **User-agency tie-up:** Solicited
- **Name and address of the Resource Person:**
Dr (Ms) Sanghamitra Mohanty
RC-ILTS-Oriya, Dept. of CSA,
Utkal University,
Bhubaneswar, Orissa.
E-mail : sanghami@rediffmail.com
Web : http://www.ilts.utkal.org

5.1.3

- **Name of the Technology Akhar :** ***The Bilingual Punjabi English Spell Checker***
- **Nature of Technology :** Knowledge tool (Software for Punjabi Language)
- **Level:** (Product / Technology / Sub-system) Product
- **Technical Description of the Technology / Product** including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product : **"Akhar - The Bilingual Spell Checker"** is the first ever spell checker for Punjabi Language and supports spell checking for Punjabi as well as English Language under the windows environment. It has MS-Word compatible features. The Bilingual Spell Checker automatically detects the wrong words and suggests the possible correct spellings. The spell checker successfully solves the problem of time consuming proof reading for the Punjabi text. The Spell Checking in Punjabi language makes use of Gurmukhi Script as the base of Punjabi language.

  Akhar makes use of a strong database of both Punjabi and English language words. The database for Punjabi makes use of Harkirat Singh's *Shabad Jor Kosh* as the base and additionally has words taken from the dictionaries published by Punjabi University, Patiala and the State Language Department and the corpus developed by CIIL, Mysore. The database for Punjabi words has been incorporated after consultation with the Punjabi linguists. In the absence of standardization of spellings in Punjabi language for some of the commonly occurring words, which have more than one spelling, all the commonly used spellings of a word have been incorporated in the database of Akhar after consultation with the Punjabi Linguists. A list of similar sounding words and consonants was compiled and a suggestion list using this knowledge was generated based on the reverse edit distance. It is found that the right suggestion is generally generated at top of the suggestion list as the default suggestion in majority of the cases. In such a case, the user only has to confirm the default suggestion and proceed with the next error. Otherwise, the user needs to scroll through a list of suggestions and pick one as the right one. The spell checker supports the text typed in any of the popular Punjabi fonts as well as ISCII encoded files. It provides a number of features that make the use of Punjabi Language on a computer easy and provides a number of options to increase the efficiency of the user while spell checking.

  Features:

  1) Very simple user interfere

  2) Bilingual Spell Checker for Punjabi and English.

  3) Support for .TXT, .DOC, .RTF and .HTML etc formats.

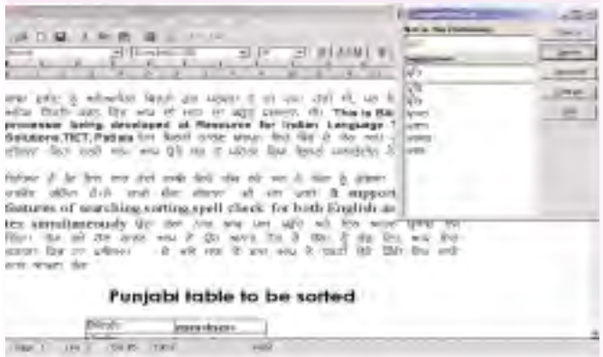
  4) Extensive help at various levels to make it easy for the user to learn.

  Platform: IBM Compatible PC with 486 or Pentium CPU

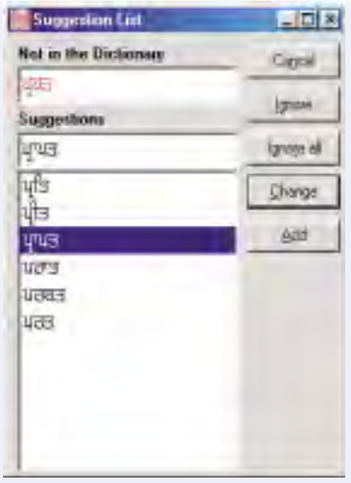
  Operating System: Windows 95,98,2000, XP

  Programming Interface: Delphi

- **Representative Snapshot / screenshot of the Technology / Product :** Screenshot of Akhar : Screenshot of Suggestion List in Akhar



- **Scalability / Portability / Expandability:** Portable among any of the Windows environment Win 95, 98, Window-ME, Win-2000, Win-NT.



- **Readiness of Transfer of Technology (ToT) :** Yes
- **Availability of documentation :** Yes
- **Testing of the Product / Technology :** Testing done
- **IPR / Open-source :** IPR obtained
- **Potential Beneficiaries :** Typists, Linguists, Students, Teachers, Punjabi Learners, Educationists, Office Personnel, Corpus Developers etc.
- **User–agency tie-up :** Modular Infotech Pvt. Ltd., Pune
- **Name and address of the Resource Person:**
  Dr. R.K. Sharma,
  Coordinator,
  Resource Centre for Indian Languages Technology Solutions – Punjabi (RCILTS - Punjabi),
  School of Mathematics and Computer Applications (SMCA),
  T.I.E.T, Patiala - 147004
  Punjab (India).
  E-mail : rksharma@tiet.ac.in
  Web : http://punjabirc.tiet.ac.in/recentrei.asp

5.1.4

- **Name of the Technology :** ***Marathi Spell Checker***
- **Nature of Technology :** Knowledge tool
- **Level:** (Product / Technology / Sub-system): Sub-system
- **Technical Description of the Technology / Product including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product:** Under this ongoing project, a stand-alone spellchecker is being built for Marathi. The spellchecker will be available to spell check documents in a given Encoding.

  From the CIIL (Central Institute of Indian Language) corpus, 12,886 distinct words have been listed. Similarly other Marathi texts at the center are being used to build a basic dictionary. A morphological analysis is being carried out on the collection of words. For example, an automatic grouping algorithm identified 3,975 groups out of 12,886 distinct words. First word is usually the root word. Thus, there are approximately 4000 root words from Marathi corpus. A manual proof reading will be done on these results. The morphology will be enriched.

  A motivation behind the stand-alone spellchecker is that it can be used without an editor through a packaged interface, or it can be integrated with other compatible applications such as OCR.
- **Representative Snapshot / screenshot of the Technology / Product:** Not available
- **Scalability / Portability / Expandability :** Scalable, Portable and Expandable
- **Readiness of Transfer of Technology (ToT) :** Ready in 6 months
- **Availability of documentation :** Documentations are available
- **Testing of the Product / Technology :** Beingsubmitted to STQC for testing
- **IPR / Open-source :** open source, subject to permission from MICT

- **Potential beneficiaries :** The potential beneficiaries are Indian language content creators, interface designers, search engine industry, translators and translation industry, text-mining industry, linguists and typists

- **Name and address of the Resource Persons:**

Prof. Pushpak Bhattacharyya and
Prof. Rushikesh Joshi
Department of Computer Science and Engineering
Indian Institute of Technology
Mumbai- 400 076 India.
E-mail : {pb, rkj}@cse.iitb.ac.in
Web : http://www.cfilt.iitb.ac.in

5.1.5

- **Name of the Technology :** ***Assamese Spell Checker.***

- **Nature of Technology :** Knowledge tool (Tool for local language incorporation in office automation.)

- **Level:** (Product / Technology / Sub-system): Product. Technical Description of the Technology / Product including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product:

- Technical Description and Algorithm used

The development of a Spell Checker for Assamese has been undertaken at the Resource Centre, and code has been developed in Perl. The Spell Checker exists in the form of separate modules for error detection and correction, as well as a stand-alone system in which all the spell checking routines have been integrated. The strategy used is described below.

Non-word detection.

The *non-words* are detected by looking up document words in a dictionary of valid words. The dictionary used is actually a word list of around 72000 Assamese words. A hash table has been used as a lexical lookup data structure. The performance of a dictionary using a hash table is quite adequate, even for complex access patterns. Perl has an efficient implementation of a hash table data structure, which has been used in the non-word detection module.

Isolated-word error correction.

A three-pronged strategy has been used for generating suggestions comprising of the Soundex, Edit-distance and Morphological processing methods.

Soundex method

This method maps every word into a key (code) so that similarly spelled words have similar keys. A Soundex encoding scheme for Assamese has been designed based on the encoding scheme for English which comprises of a set of rules for encoding words and 14 numerical codes. The

Soundex code of the misspelt word is computed, and the dictionary is searched for words, which have similar codes. Some example are given below (Table 2).

Edit-distance method

Candidate suggestions are obtained by 'Damerau's error reversal' method. The four well known error-inducing edit actions are the insertion of a superfluous letter, deletion of a letter, transposition of two adjacent letters and substitution of one letter for another. In 'Damerau's error reversal' each edit action is applied to the misspelt string and a set of strings is first generated. These are checked in the dictionary to see which of them are valid words to finally produce the suggestions.

Basic block diagram

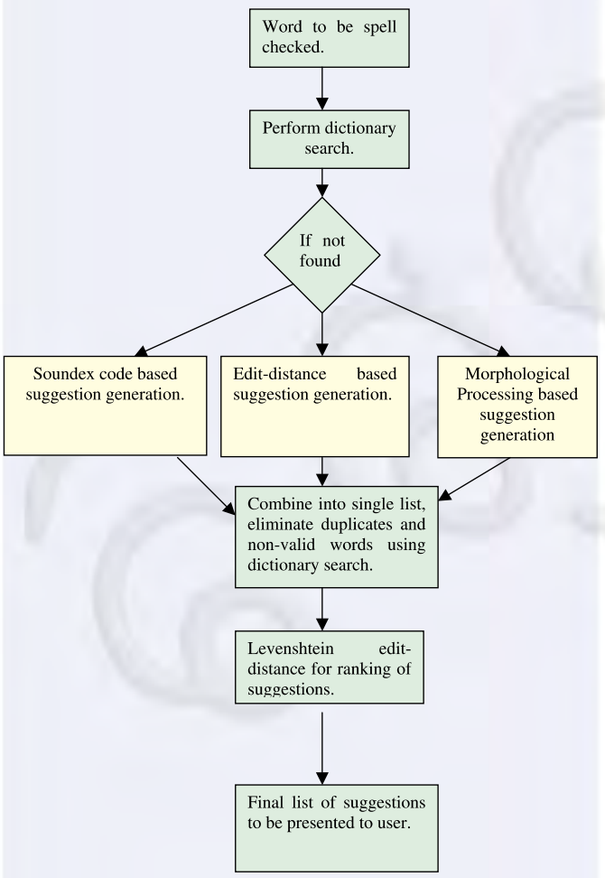


Figure 1. Strategy followed by the Assamese Spell Checker.

Word	Soundex code
মৰম meaning affection	ম৩৫
স্বাধীন meaning independent	স্বা৩৫
ফঁহিয়ালেহেঁতেন meaning would have analysed	ফ৭৯৪৭৩৫

Morphological processing Morphological analysis is performed to extract the root word from the misspelt word as also a list of valid affixes that can be attached to that root word. By attaching the affixes that closely match those of the misspelt word to the root word, a list of suggestions is generated.

A module for ranking of suggestions, based on Minimum edit-distance methods has also been developed.

O/S used

Currently tested on Windows '98.

Front-end / user interface

The user interface for the Assamese Spell Checker has been built using Perl/Tk. A runtime snapshot of the Assamese Spell Checker is shown in Fig. 2.

Specifications

Features of the stand-alone Spell checker

- GUI developed using Perl/Tk, with simple text editing facilities.
- Currently supports C-DAC's AS-TTDurga font.
- Assamese text files can be loaded onto the GUI, edited, and saved.
- Misspellings can be marked by clicking the 'Det' ('Detect') and then the 'Show' button.
- Selecting the misspelt word and clicking 'Sug'('Suggest') button can generate suggestions.

- A facility to add new words to the dictionary exists.
- 'Select All/Unselect All' option available in the 'Edit' menu on right clicking on the text, for selecting/deselecting entire text.
- 'Copy'/'Cut' and 'Paste operations possible for text within the GUI window.
- Text can be copied/cut from an I-leap document and pasted onto the GUI window.

- **Representative snapshot / screenshot of the Technology / Product:**

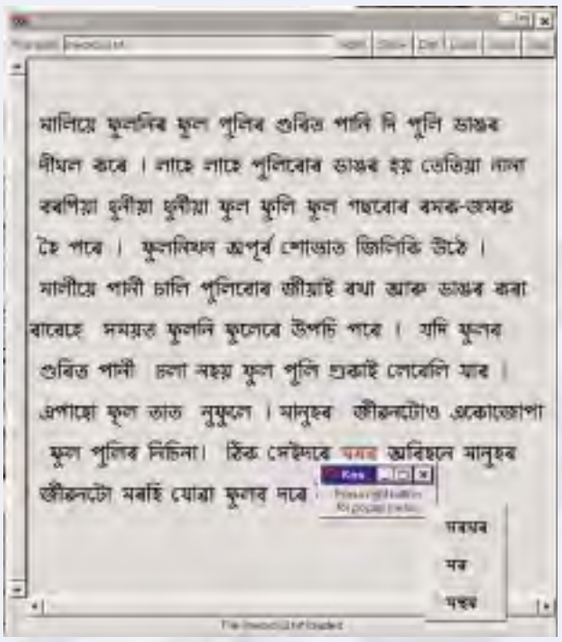


Figure 2. The Assamese spell checker.

- **Scalability / Portability / Expandability:** Portable.
- **Readiness of Transfer of Technology (ToT):** Installation guidelines and User guides to use and configure the system are under preparation.
- **Availability of documentation :** Under preparation.
- **Testing of the Product / Technology :** In progress.
- **IPR / Open-source :** Open Source.
- **Potential beneficiaries :** DTP Operators, State Government Employees.
- **User-agency tie-up :** Solicited
- **Name and address of the Resource Person:**

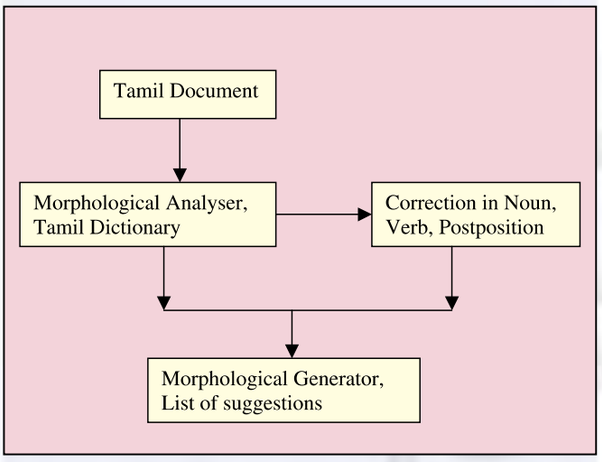
M.Das, moni@iitg.ernet.in
S.B.Nair, sbnair@iitg.ernet.in
Department of Computer Science & Engineering,
Indian Institute of Technology Guwahati
Assam – 781039 (India)
Web : http://www.iitg.ernet.in/rcilts

5.1.6

- **Name of the Technology :** ***Annam (Tamil Spell Checker)***
- **Nature of Technology :** Linguistic Tool
- **Level :** Product
- **Technical Description of the Technology Description :** Tamil Spell Checker is a tool used to check the spelling of Tamil words. It provides possible suggestions for erroneous words. User has the provision to select the suggestion among the list, ignore the suggestion or add the particular word to the dictionary.

Block Diagram

Tamil Spell Checker



Spell Checking in Noun & verb modules

These modules extract the root word from the given word (noun/verb) with the help of Morphological Analyser and the root word is checked in the dictionary and if found, the word is termed as correct word. Otherwise, the correction process is activated. The correction process includes error handling and suggestion generation modules.

After finding the type of error, the right form of suffixes; nouns or verbs are given as input to suggestion generation module. With the help of Morphological Generator, the correct word is generated. This module also handles the operation like select, change or ignore the suggested word and adding the word to the dictionary.

Errors handled in Noun module

- Case marker, plural marker checking
- Postposition checking
- Adjective checking
- Adjacent key error checking
- Similar sound letter checking Errors handled in Noun module
- Person, number and gender marker checking
- Adverb checking
- Adjacent key error checking
- Similar sound checking

4C. User Interface

The user interface helps to check for spelling of the word. 'Open' a file for spell check, 'Spell check', 'Suggestion' for wrongly spelt words, 'Cancel' and 'Exit' buttons are provided in the interface. The suggestion window is displayed when the user selects the suggestion button. A suggestion list is provided for the wrongly spelt words. This interface has 'Change', 'Ignore', 'Add' it to the dictionary and 'Exit' buttons.

4D. Specifications:

DEVELOPED ENVIRONMENT

Platform : Windows 2000

Software Components : JDK 1.3.0

HARDWARE CONFIGURATION

Processor : Intel Pentium PIII

RAM : 128 MB

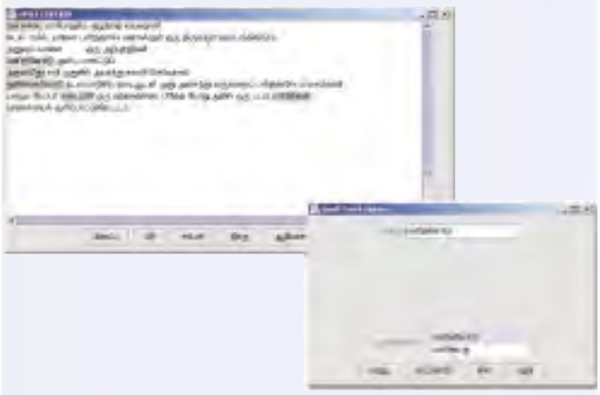
Hard Disk : 20.0 GB

4E. Requirements

- JRE 1.3
- Any Tamil keyboard driver (Font encoding in TAB) to type Tamil.

- Code converter for other font encoding is provided.

- **Representative Snapshots:**



6A. Scalability : Yes

The Spell checker can be improved to handle compound word errors.

6B. Portability : Yes

- **Readiness of Transfer of Technology :** Yes
- **Availability of Documentation :** User and Technical manuals are available.
- **Testing :** The Tamil Spell Checker is tested with large Tamil corpus.
- **IPR / Open Source :** IPR
- **Potential Beneficiaries :** Public
- **User Agency Tie up :** RCILTS-Tamil has tied up with MODULAR INFOTECT, Pune.
- **Name and address of the Resource Person:**
Dr. T.V. Geetha,
Dr. Ranjani Parthasarathi,
Resource Centre for Indian Language Technology Solutions,
School of Computer Science and Engineering,
Anna University, Chennai – 600025, India.
E-mail : rp@annauniv.edu
Web : http://ns.annauniv.edu

5.1.7

- **Name of the Technology :** Malayalam Spell Checker (NERPADAM™)
- **Nature of Technology :** Tools
- **Level :** Sub system
- **Technical Description :** NERPADAM™ is a software subsystem that can be integrated with Microsoft word as a macro or the Malayalam editor, stylepad developed by CDAC, Thiruvananthapuram, to check the spelling of words in a Malayalam text file. While running as a macro in word, it functions as an offline spell checker in the sense that one can use this software with a previously typed text file only. Both off line and online checking are possible when it is integrated with the text editor. It generates suggestions for wrongly spelt words.

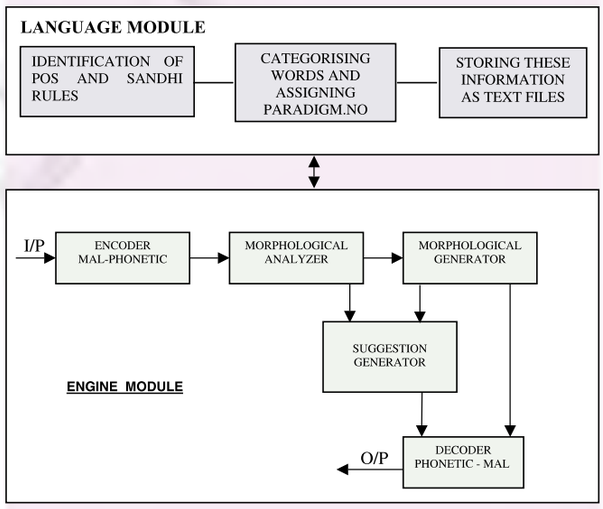


Figure 3.1 Block Diagram of Spell Checker

The system adapts a rule cum dictionary-based approach for spell checking. It incorporates a fully developed Morphological Analyser for Malayalam. This module splits the input word into root word, suffixes, post positions etc. and checks the validity of each using the rule database. Finally it will check the dictionary to find whether the root word is present in the dictionary. If anything goes wrong in this checking it is detected as an error and the error word is reprocessed to get 3 to 4 valid words, which are displayed as suggestion. The user can add new words into a

personalised data base file, which can be added to the dictionary if required.

- **Portability/Expandability/scalability:** Presently we are working on increasing the accuracy of the suggestion generation.
- **Readiness of Transfer of Technology (ToT):** The Product is ready for Transfer of Technology.
- **Availability of documentation:** Design /TOT documents available.
- **Testing of the Product/Technology:** Third party testing not done.
- **Potential beneficiaries:** Any body doing Malayalam word processing jobs.
- **User-agency tie-up:** Solicited
- **Name and address of the Resource Person:**
R. Ravindra Kumar; K.G.Sulochana
Additional Director & Coordinator;Dy. Director
RCILTS- Malayalam
C-DAC, Thiruvananthapuram
Email: ravi@erdcitvm.org.
Email: sulochana@erdcitvm.org.
Web : http://www.malayalamresourcentre.org

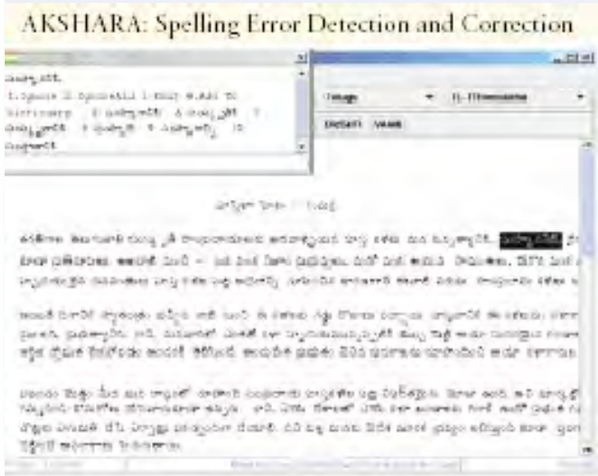
5.1.8

- **Name of the Technology : AKSHARA: Telugu Spell-Checker**
- **Nature of the Technology :** Knowledge tool
- **Level :** Technology Transferred to M/S Modular Infotech .
- **Technical Description :** Perhaps the best and most thoroughly worked out morphological analyzer for Telugu is the one we have developed at University of Hyderabad over the past 10 years or so. The system has been tested on large scale corpora and enhancements and refinements have been going on for years. During this project, a thorough re-engineering work was taken up and a new version was developed. The new version is far simpler, more transparent, portable, well documented, and conforms to standards.

  Research work has also been taken up on developing stemming algorithms for Telugu. A pure corpus based statistical stemming algorithm has been developed. The performance of this stemmer for the spell checking application has been studied in various combinations with dictionary and morphology based approaches.
- **Error Pattern Analysis:** Large scale spelling error data has been obtained from our 10 Million word Telugu corpus. The raw corpus as it was typed has been compared with the final version after three levels of proof reading and certification by qualified and experienced proof readers. A number of tools have been developed to prepare such a data. Quantitative study of spelling error patterns in Telugu is being conducted. This will help us to build better spell checkers in future.
- **Syllable level statistics for spell checking:** Since words are large and complex and hence too numerous in Telugu and proper morphological analysis is difficult, it would be useful to perform studies at lower levels of linguistic units. The syllable level is a natural choice since writing in Indian languages is primarily syllabic in nature. n-gram models have been built at syllable level.

HMM models have also been built. These models can be used to detect spelling errors and to rank the suggestions for correcting a given word.

- **Representative Screenshot**



- **Operating System used :** LINUX
- **Scalability / Portability / Expandability :** Available
- **Availability of documentation:** Yes
- **Testing of the Product / Technology :** Testing has been done
- **IPR / Open Source :** IPR lies with University of Hyderabad and Dept of Information Technology.
- **Potentical beneficiaries :** Telugu word processing software developers.
- **User agency tie-up :** Negotiation with Modular Infotech.
- **Name and address of the Resource Person :**

Dr. K. Narayanamoorthy
Department of CIS, University of Hyderabad
Hyderabad-500046
A.P. India
E-mail : knmcs@uohyd.ernet.in
Web : http://www.languagetechnology.ac.in

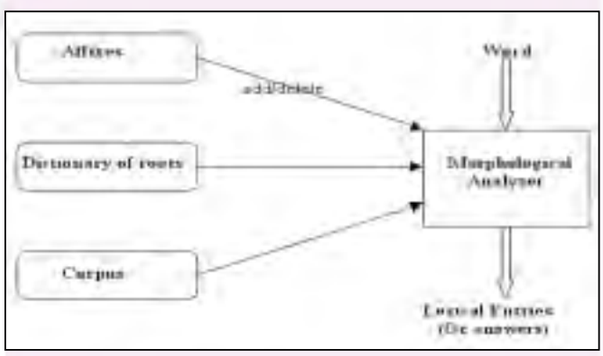
5.2 Morph Analyzer

5.2.1

- **Name of the Technology :** ***Assamese and Manipuri Morphological Analyzers.***
- **Nature of Technology** : Knowledge tool (Aid to machine translation system and Spell Checker.)
- **Level: (Product / Technology / Sub-system):** Technology
- **Technical Description of the Technology / Product including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product:**
- **Assamese Morphological Analyzer:** The analyzer has been developed for use with the Spell checker and the Machine translation systems.

  Technical Description Stemming technique forms the base of the Assamese morphological analyzer. In the above technique, affixes are added/deleted according to the linguistic rules. The derived words are verified with the existing Corpus/Dictionary to treat as valid words.

  Block Diagram



Features

a) Currently works with Twenty linguistic rules

b) More rules can be added without alteration in code.

c) Modules are available in the form of API's for customization.