



Sanchay

**An NLP Platform for Indian
Languages**

**Anil Kumar Singh <anil@research.iiit.ac.in>
Language Technologies Research Centre
International Institute of Information Technology
Hyderabad, India**



Sanchay: A collection of APIs and tools for NLP

- Focuses on Indian languages
- Includes GUI based interfaces for annotation
 - Like Syntactic Annotation Interface
 - Parallel markup interface



Some Features

- Object oriented
- Reusable and flexible/customizable components
- Every component an API
- No commitment to one specific architecture or theoretical framework
- GUIs for as many things as possible
- Open source



Major Components

- APIs for language resources
- Data structures for usual kinds of data
- Basic GUI components (table, tree, etc.)
- Annotation interfaces
- Sentence alignment tool
- Language and encoding identification tool
- A multi-purpose editor for NLP and Indian languages
- A computational phonetic model for Indian language scripts




Resources

- Atomic resources
 - Raw corpus
- Aggregate resources
 - Parallel corpus



Some GUI Components

- Enhanced Tree
- Enhanced Table
- Tree Viewer



Implementation of NLP Algorithms/Techniques

- Sentence alignment
- Language and encoding identification
- N-Gram modelling
- Dictionary Trie



Data Components

- Trees representing SSF, XML, etc.
- A generalized table
- Properties manager



Utilities

- File splitter
 - For different kinds of corpora
- Find/replace/extract
 - Uses regular expressions
 - Can work in batch mode
- Preprocessing of raw text
- Many others



Support for Indian Languages

- Works without installing fonts
- Many major Indian languages supported
- Uses a collection of free fonts
- Font listing according to the language and encoding



Interfacing with Other NLP Platforms/Libraries

- Not yet done, but planned:
 - GATE
 - OpenNLP
 - Mallet
- Tried a Hindi POS tagger using OpenNLP MaxEnt package



Future Work

- Generalized support for XML based annotation
- Applications like spell checker for Indian languages