4.1 Corpora (Parallel Corpora)

4.1.1

- **Name of the Technology :** *GyanNidhi – 1 Million Pages Parallel Multilingual Corpus*

- **Nature of Technology :** Knowledge Resource (Content creation and tools development)

- **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 :** Gyan-Nidhi : Multilingual Parallel Content Creation targeting 'One Million Pages' for the Knowledge Repository. Gyannidhi is a collection of multilingual parallel text in Indian languages. The project envisages on the collection of 12 Indian languages (Hindi, Punjabi, Gujarati, Marathi, Tamil, Telugu, Kannada, Nepali, Oriya, Malayalam, Bangla, Assamese) and English.
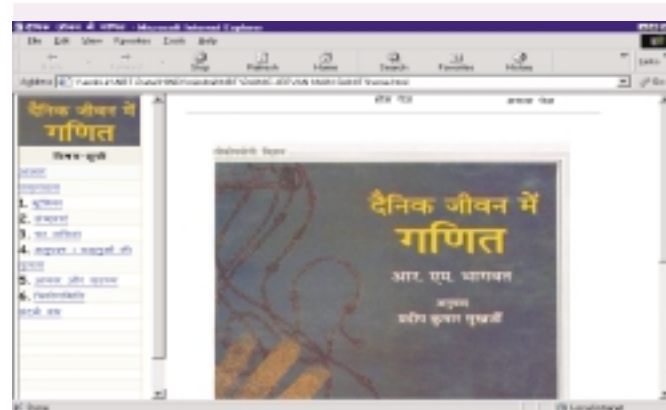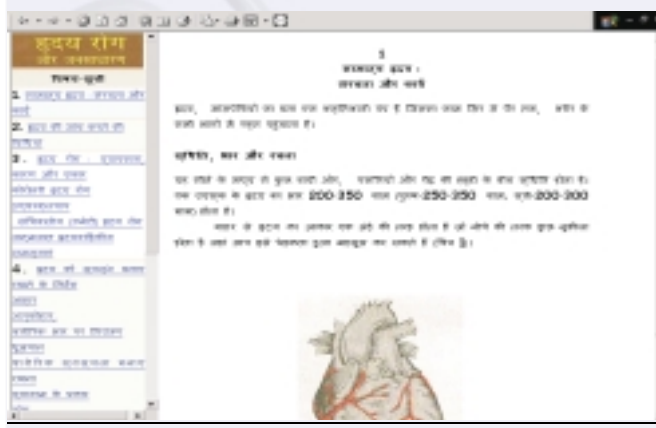
Features:

1) State of the art GUI.

2) Corpus collected from various domains.

3) Text in Unicode (International Standard).
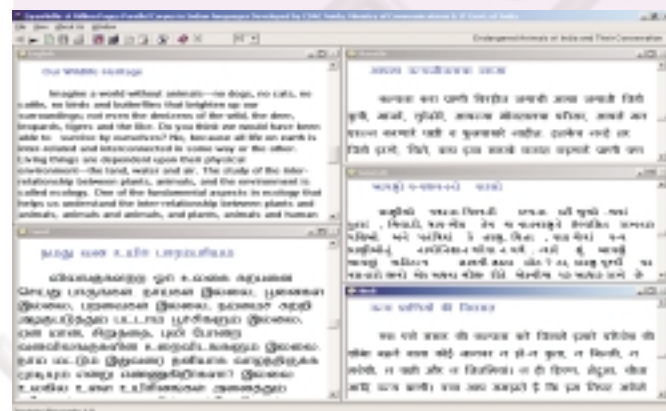
4) Corpus management tool.

Text Encoding: Unicode/ISFOC

User Interface: Visual Basic

- **Representative Snapshot / screenshot of the Technology / Product:**
  Screenshots:

Screenshot of Html format book content

Screenshots of the corpus management tool

- **Scalability / Portability / Expandability:** Content portable among any of the Operating system which supports which supports Unicode.

- **Readiness of Transfer of Technology (ToT) :** Yes

- **Availability of documentation :** Yes

- **Testing of the Product / Technology :** Testing done

- **IPR / Open-source:** IPR lies with Dept. of IT

- **Potential Beneficiaries :** Researchers, Linguists, Translators, Typists, Educationists, Students etc.

- **User–agency tie-up :** Like minded research institutions (solicited)

- **Name and address of the Resource Person:**
  V N Shukla, Dir(Spl App) C-DAC,
  C – 56/1, Sector 62,
  Noida – 201 307,
  Phone : 0120 – 2402551...558
  E-mail : vnshukla@cdacnoida.com
  Web : http://www.cdacindia.com

Knowledge Resources