Multilingual Book Reader: Transliteration, Word-to-Word Translation and Full-text Translation

Prashanth Balajapally
Sai Spurthi Institute of Technology
prashant.bala@gmail.com

Phanindra Pydimarri
Institute Of Aeronautical Engineering
phanindra.pydimarri@gmail.com

Madhavi Ganapathiraju
School of Computer Science
Carnegie Mellon University
madhavi@cs.cmu.edu

N. Balakrishnan
Supercomputer Education and Research Center
Indian Institute of Science
balki@serc.iisc.ernet.in

Raj Reddy
School of Computer Science
Carnegie Mellon University
rr@cmu.edu

Abstract:

India being a multilingual nation, with 22 recognised official languages, also has literature in all these languages; they find representation in the Digital Library of India (DLI) which holds over 120,000 books. DLI has driven the creation of a large number of applications to process and present the Indian language content. In this paper, we present the creation of a multilingual book reader interface for DLI that supports transliteration and "good enough translation" features making it possible for readers to read a book that is written in another language.

Introduction

India is a nation with pluralistic culture, a large number of cultures, ethnicities, languages and religions coexisting with each other. While the culture and faith unify the country under one umbrella either by similarity or by tolerance, the language is what separates them. In the 1951 census, the first census after India attained independence, 845 languages (dialects) were identified, of which 60 were spoken by at least 100,000 people each. The Indian constitution identifies 22 languages, of which six languages (Hindi, Telugu, Tamil, Bengali, Marathi and Gujarati) are spoken by at least 50 million people within the boundaries of the country—there are a large number of them living outside the country. Although the Indian languages were identified as belonging only to four different language families, namely, the Austric, Dravidian, Tibeto-Burman, and Indo-Aryan, the language spoken by one person is rarely understood by a person speaking another language; this does not however rule out bilingualism of a large number of people, especially those who migrate from one state to another, where they speak the mother tongue at home and can usually follow the dominant language of the new state. For example, Telugu speakers are found in good numbers in Karnataka (3,325,062), Maharashtra (1,122,332), Orissa (665001), and Tamil Nadu (3,975,561); about 10% of Telugu speakers live outside of the Telugu territory, according to an old 1901 estimate; this number would be much larger today. Bilingualism is also found at the borders of two states, where people can usually speak languages of both the states sharing the border. Taking the example of Andhra Pradesh again, where the native language is Telugu, a large number of people speak languages of its neighbours: Kannada (519,507), Marathi (503,609), Oriya (259,947), and Tamil (753,484).

Digital Library of India is an effort to bring the advances in information technologies towards preserving the rich Indian heritage present in the form of literature, art and manuscripts, by converting them into digital form. DLI holds a collection of close to 120,000 books in Indian languages and English (Balakrishnan and Reddy 2004; Balakrishnan, Reddy, Ganapathiraju et al. 2004). Table 1 shows the number of books and pages in current holdings in various different languages. The collection consists of mathematics, science and literature just like in a physical library. These books of the DLI are available to anyone, anytime, anywhere. The goal of the OMBRÉ project is to add the dimension of any-language to it. To this end, we developed a multilingual book reader that supports automatic transliteration and word-to-word translation between Indian languages and English.

Language	Books	Pages
English	53596	19167998
French	145	30137
German	117	48170
Greek	1	539
Hindi	3248	642905
Italian	6	4660
Kannada	487	116036
Marathi	167	38954
Norwegian	5	1452
Spanish	41	8998
Sanskrit	2532	945235
Persian	1106	347282
Tamil	372	83314
Telugu	18231	3811111
Urdu	1961	442965
Other	826	261661

Table 1: The language-wise break-up of the first 82,000 books are shown. The books contain largely Indian language books and those in English. Books from foreign languages are also present, especially very old books for preservation purposes. Copyright issues are discussed elsewhere, but in short, the books scanned are those that are out of copyright (pre-1920 books), and those that are out of print to which rights are obtained from the authors/publishers.

Indian Language Technology Research under DLI

The primary goal of DLI, apart from making the books available online, is to make them available in a fully functional form. The DLI is not simply a static repository of books—it has made possible bringing home language and information processing technologies for Indian languages. DLI acts as a catalyst for Indian language technology research—advances have been made in optical character recognition (OCR), transliteration and machine translation, and Indian language lexicons, with the objective of providing the digital library user with a "good enough comprehension" of the content of the books in languages other than his own (Balakrishnan, Reddy, Ganapathiraju et al. 2005). The book reader is an integration of a number of these other components developed in the DLI.

OCR in Indian Languages- Kannada

Designing an accurate OCR in Indian languages is one of the greatest challenges in computer science. Unlike European languages, Indian languages have more than 300 characters to distinguish, a task that is an order of magnitude greater than distinguishing 26 characters. This also means that the training set needed is significantly higher for Indian languages. It is estimated that at least a ten million-word corpus would be needed in any font to recognise with acceptable accuracies in Indian languages. DLI is expected to provide such a phenomenally large amount of data for training and testing of OCRs in Indian languages. Many of the contents have been manually entered besides scanned images for this purpose. Using this extremely large repertoire of data, a Kannada OCR had been developed. The current level of accuracy that we get is around 96-97% on clean documents scanned at 400 dots per inch, and 40-50 % if the image is of bad quality. This OCR is currently being improved and also being extended to other Indian languages including Tamil.

Om transliteration: Unified representation for Indian language

There was a need for the development of a digital representation that lays a common foundation for all the Indian languages. For seamless adaptation of algorithms in language technologies, this representation must also be parsable by universal language processing tools and algorithms, such as for machine translation, information retrieval, text summarisation and statistical language modelling.

Om uses the same representation both for keyboard input and formation and digital storage. It is similar to ITRANS (Indian language Transliteration scheme: http://www.aczoom.com/itrans/), the very first widely used ASCII transliteration for Indian languages, in that it uses combinations of the English alphabet to represent Indian syllables. Om has been designed on the following principles, to enhance the usability and readability: (i) easy readability (ii) case-insensitive mapping and (iii) phonetic mapping, as much as possible. For transliteration to Indian languages, Om representation is mapped to the Indian language fonts for display or converted to any other format such as Unicode where required. When a user is not interested in installing language components, or when the user cannot read native language script the text may be read in English transliteration itself. Even in the absence of Om to native font converters, people around the globe can type and publish texts in the Om scheme which can be read and understood by many, even when they cannot read the native script. The readability criterion that is the benefit of the case-insensitive phonetic mapping proves very useful. The Om mapping tables for many Indian languages can be seen at http://www.dli.ernet.in/Om/.

Om also exploits the commonality of the alphabet (not the script) of Indian languages, and hence the representation of the same letter is the same across the many languages. This separation of storage and rendering makes it language independent across the Indian Languages. The design and creation of the Om representation and transliteration scheme forms the most important component in the book reader.

Om text editor

An integrated editor that accepts Om ASCII keystrokes as input and maps them to native fonts has been developed. The script in any one of the supported true type fonts is sent to MS Winword® for further formatting and layout options. Since the Om scheme is common to all the Indian languages, the display of the text can be converted between the supported languages by a choosing it on the menu. The text may be saved as Om (ASCII) text, native font text or in Unicode. However, for the purposes of language processing applications, the storage mechanism used is ASCII text. The Om transliteration integrated editor is available for download at http://swati.dli.ernet.in/om/ (Ganapathiraju, Balakrishnan, Balakrishnan et al. 2005).

Indian language Search Engine- Tamil Search Engine (OmSE)

If the only way of accessing the books were by metadata fields like the title, author and publisher, most of the books may forever remain unread on the DLI, because the users are usually unaware of which specific book they want to read. Hence, to access a book a good search engine is required on the front end, that supports search of metadata and full text, ensuring the QWERTY keyboard entry matches that of the phonetic text of the native language. The creation of the Om transliteration scheme makes this possible, since Om text is QWERTY keyboard enterable and is stored in plain text making it possible to search Indian language text much the same way as done for English text.

Technology for the deployment of information retrieval in Indian language has been demonstrated by the development of the OmSE searching engine using off-the-shelf open source software Greenstone search engine (Jayaraman, Sangani, Ganapathiraju et al. 2004). Tamil documents stored in the ASCII representation of Om have been built and are directly available for indexing, searching and retrieval without any modifications to the text-handling modules of the search engine. At the time of display, the retrieved text, in addition being made available in readable English transliteration, is also converted to native Tamil script and displayed. The front-end of the search engine is the user side, having a graphical user interface, which prompts the user to type in the search query in Om transliteration format. The query typed by the user is also displayed in Tamil font for the user to make corrections, if required, while entering the keyword in Om Transliteration format. Links to the retrieved results are then displayed in text format.

Machine Translation

Example Based Machine Translation (EBMT) is basically translation by analogy. An EBMT system requires a set of sentences in the source language and their corresponding translation in the target language. A bilingual dictionary comprising of sentence-dictionary, phrases-dictionary, words-dictionary and phonetic-dictionary is used for the machine translation. Each of the above dictionaries contains parallel corpora of sentence, phrases and words, and phonetic mappings of words in their respective files. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to occur again. A sentence may be seen as a combination of phrases. To translate, each sentence is divided into its constituent phrases and words, and these smaller units are translated by looking up in the

sentence, phrase and word dictionaries. For words whose translation is not found, at least their phonetic translation (transliteration) is shown in the target language.

Example Based Machine Translation (EBMT) has a set of 75000 most commonly spoken sentences that are originally available in English. These sentences have been manually translated into three of the target Indian languages, namely Hindi, Kannada and Tamil. Bilingual word and phrase dictionaries between these target languages and English of over 25,000 and 18,000 entries respectively were also created manually. Rules of translation have been created, that allow substitution of a given noun with another noun, a verb with another verb and so on, without the need to enter every combination separately in the phrase database. This has been observed to improve the results dramatically. The phrase translations and phrasal rules play a significant role in this translation system. The advantage of this simple “good-enough translation” system is that its performance can be improved almost linearly with the increasing corpus and rule base, and especially for translating between 2 Indian languages for informal usage; the good enough translation is useful since the languages have a common root and hence share a large number of words across the different languages.

The web-enabled version of Example Based Machine Translation is available at http://bharani.dli.ernet.in/ebmt/. The current machine translation system supports the following language-pair translation:

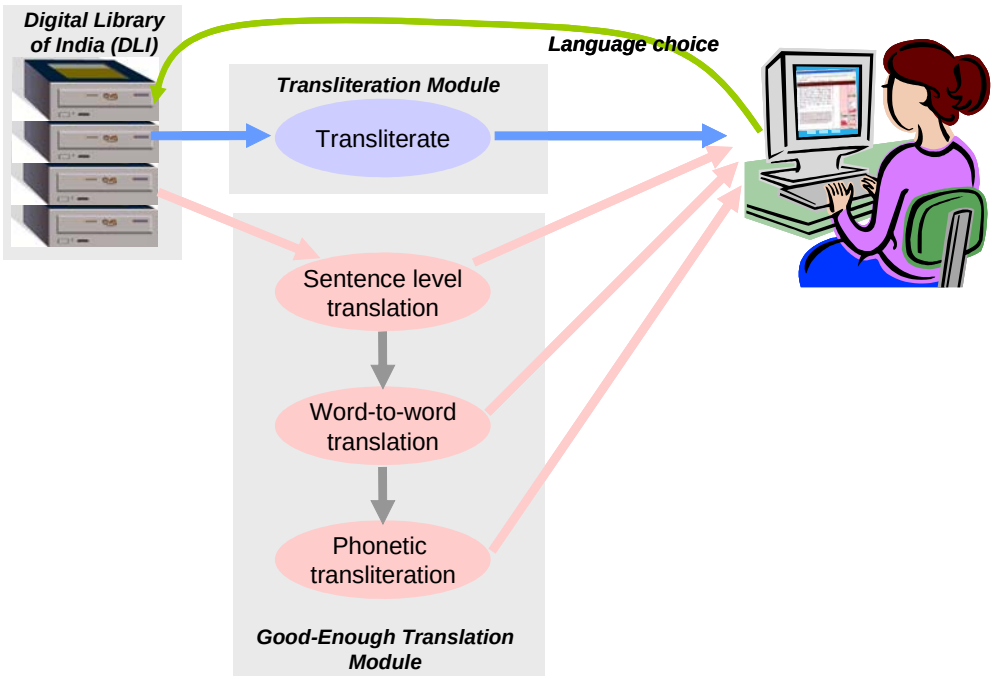
1. English to Hindi
2. English to Kannada
3. English to Tamil
4. Kannada to Tamil.

Multilingual book reader interface

The books on the Digital Library of India are available to anyone, anytime, anywhere. The book reader adds to it, the dimension of any-language. While Om transliteration helps one to read the text of one language with a script of another, it does not provide any translation. Due to the grammatical and etymological similarity amongst Indian languages, and their phonetic similarity, Om effectively goes a step beyond mere transliteration. Understanding of the text in a different language can be improved further by a simple technique of merely translating only some frequently occurring words from corpora. This has been the motivation behind the development of a multilingual book reader that supports automatic transliteration and word-to-word translation between Indian languages and even English. A dictionary, with the objective of providing the digital library user with a “good enough comprehension” of the content of the books in languages other than his own, has been built. The cross-lingual dictionary currently has six Indian Languages (Hindi, Telugu, Assamese, Tamil, Kannada and Malayalam) besides most of the European languages.

When presented with electronic text in any Indian language, the book-reader allows text to be transliterated into any one of the many Indian languages. This is made possible with the Om transliteration scheme that is discussed above. This allows the user to read for example, Hindi text in Telugu font. With the help of a Universal Dictionary, a word-to-word look up table translation is made on the Indian language text between any pair of the many Indian languages supported by the Universal Dictionary. When a word is not found in the look-up table, only its transliteration is displayed. While Indian languages are phonetic languages, English is not phonetic. In order to display English words in native language where required, a pronunciation dictionary is used.

These features provided by the interface are desirable not only to the readers who can understand but not read their own language, but also to those who desire to obtain at least a crude translation of the book to their desired language. The book reader performs the functions involving transliteration independently, while also connecting to the example-based machine translation system on the backend for full-text translation. The reader, while especially suited to a multilingual country like India, is also extendable to any other digital library, where the resources of translation and transliteration are available at large. The multilingual book-reader presents novel features that improve the usability and reach of any digital library.



OMBRE: User selects a book to read on the Digital Library of India, and if interested, chooses multilingual option. Under this option, she can choose the language in which she would like to read the content, and also a choice of transliteration, word-to-word translation, or full-text translation. These features are available when the book is available as plain text after optical character recognition. Transliteration is possible between any-to-any Indian languages, since the plain text is represented in the Om transliteration scheme. The option of word-to-word translation is made possible by the Universal Dictionary. As shown in figure above, if this fails (if the word is not found in UD), then that word is shown only as a transliteration and not as a translation. Similarly, sentence level translation relies on example based machine translation system developed under DLI. Again, the philosophy of "good enough comprehension" reverts to the lower level translation (word-to-word), if the sentence level translation fails. Thus, the user is presented with parts of texts that are translated and parts that are translated at word level, and where that is not possible as well, with just a pronunciation option, allowing her to get a "good-enough comprehension" of the content of the book.

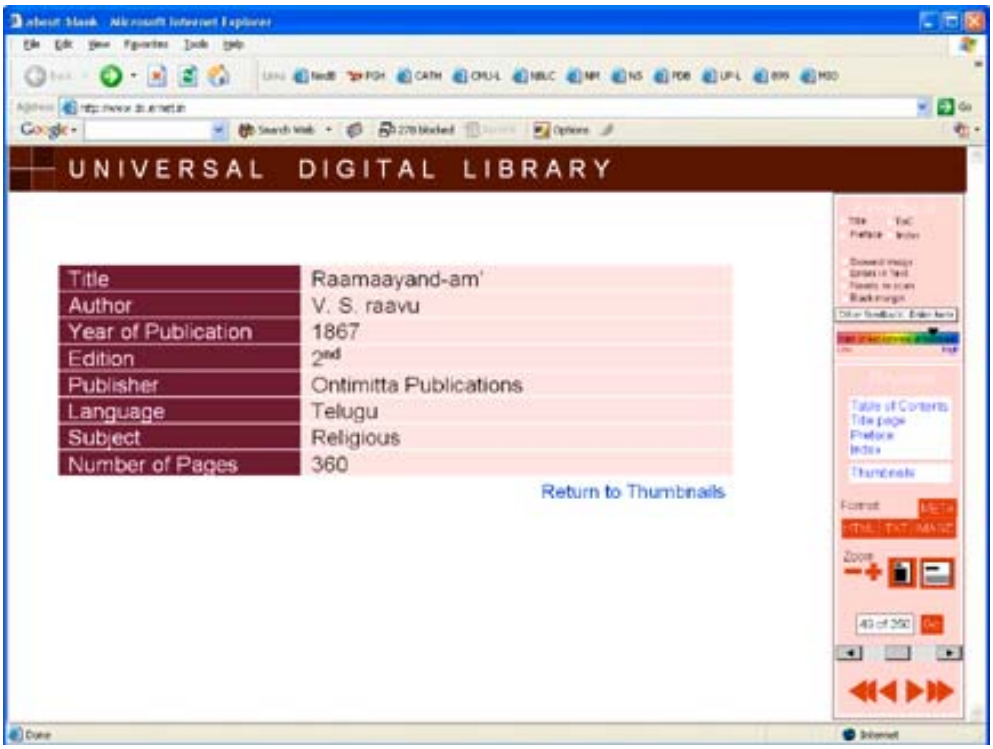
When a book is retrieved by browsing or searching on the DLI (see Snapshot 1), the user can choose either to view the book in a normal fashion (Snapshot 2) or choose to user the multilingual interface to read the book. When the latter is chosen, the user is also given a choice of languages of which he can choose one as the target language in which to read the book (Snapshot 3). The user makes a choice of the target language and can then proceed to use the multilingual tools, namely transliteration, word-to-word translation and full text translation, which are described in this section.

Transliteration

When this option is chosen, the viewing panel of the book is divided into two parts one below the other. In the top panel, the text in the original language is shown. In the bottom

panel, text transliterated to the chosen target language is shown (see Snapshots 4, 5 and 5 for Telugu text transliterated to Tamil, Hindi and English; one can read the English text aloud and what will be heard is actually the Telugu language, meaning that only the written form has been changed). For most Indian users this is sufficient, as this means being able to understand the text in their mother tongue, even though they did not learn to read it at school, or reading the book in a language that they already understand. Thus, these users may not look beyond this transliteration. Note that the text transliterated to English is readable in quite a natural way, and it sounds even more natural when transliterated to another Indian language.

Snapshot 1 ↓



Word-Word translation

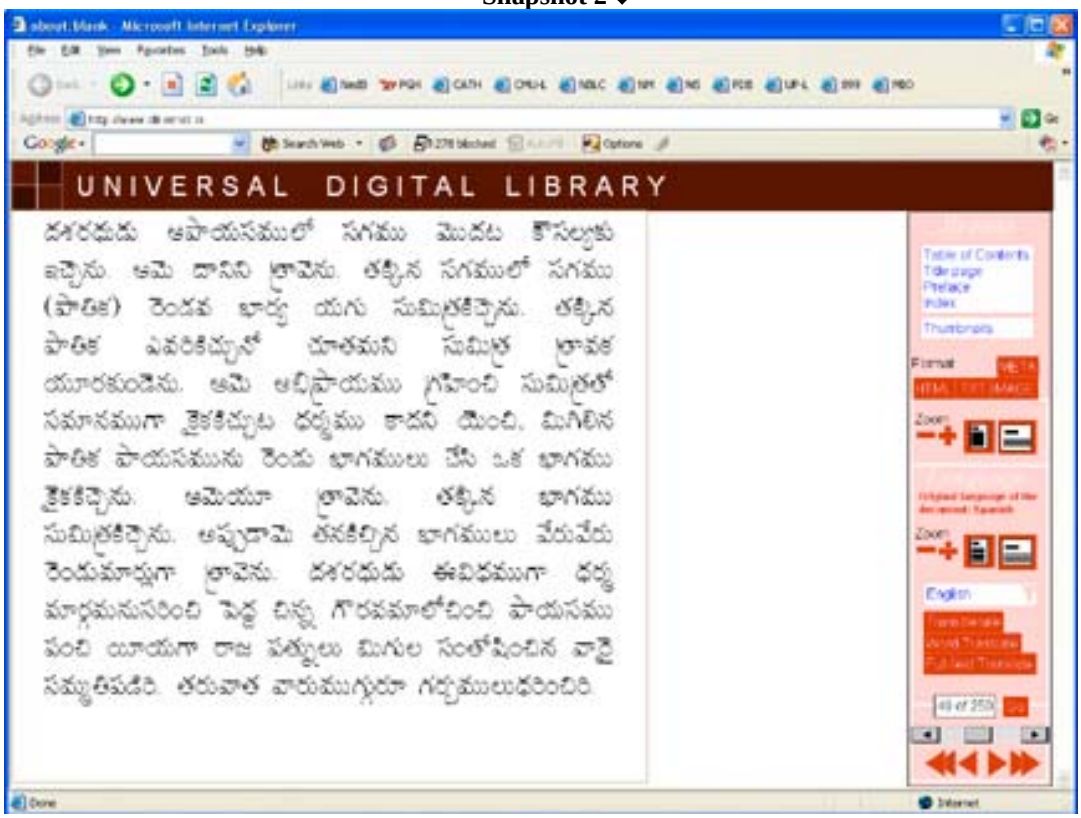
When the user is not sufficiently familiar with the source language, he can choose to open the book in another window and choose to perform a word-to-word translation. When this option is chosen, word-to-word meanings are matched from a lexicon, and displayed in the bottom panel. To indicate the correspondence of words from top and bottom panels, alternating colours of blue, green and red are used to colour the units of words for which meaning has been found in the lexicon (see Snapshot 7). Where an exact match is not found in the lexicon for a word, it is stemmed using a Porter stemmer (Porter 1980). The entries of source language in the lexicon are also stemmed and stored on the web server. So the word that needs to be translated from the book is matched with the stemmed words of the lexicon. If a match is found, its meaning is displayed, just as the meaning of a whole word would be displayed, with matching colour between source and target language panels. If no match is found, the word is displayed as it is, but in a transliterated form.

Full-text translation

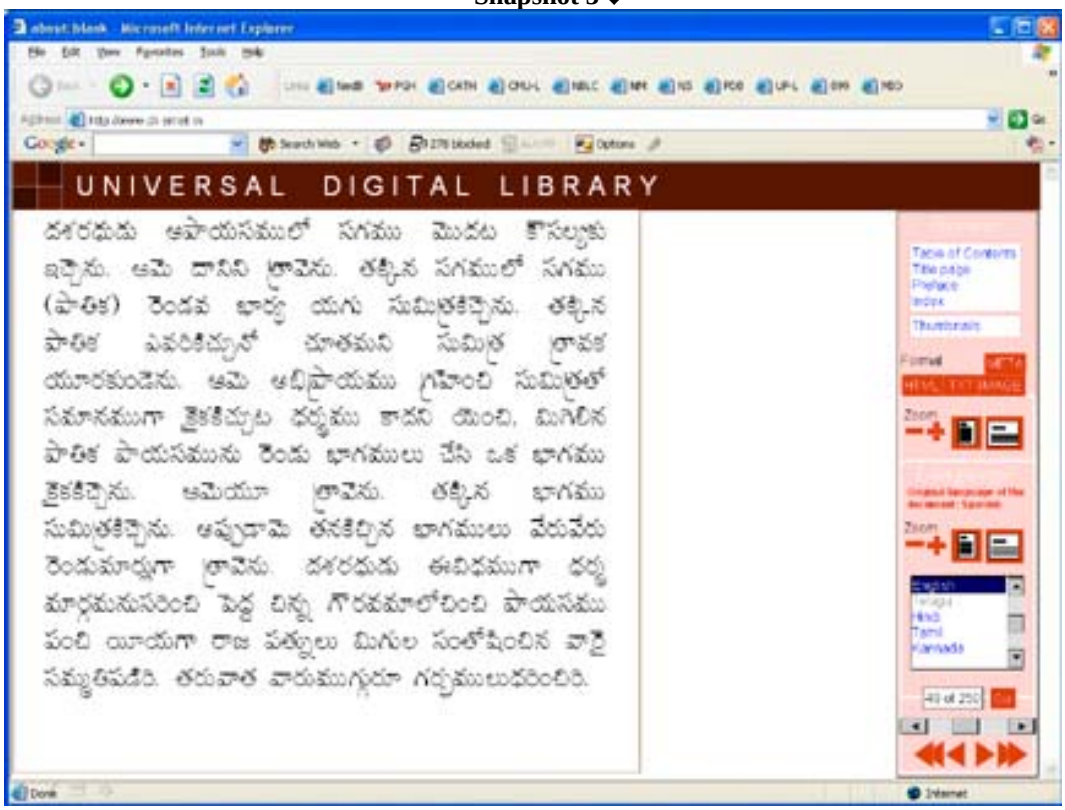
To perform this, the text in the book is broken into sentences—since Om allows the use of sentence capitalisation and “full-stop” at end of a sentence, the sentence separation rules are

same as those in English. Each sentence is then sent to the EBMT engine for translation, and the output is displayed in the target-language bottom panel (see Snapshot 8).

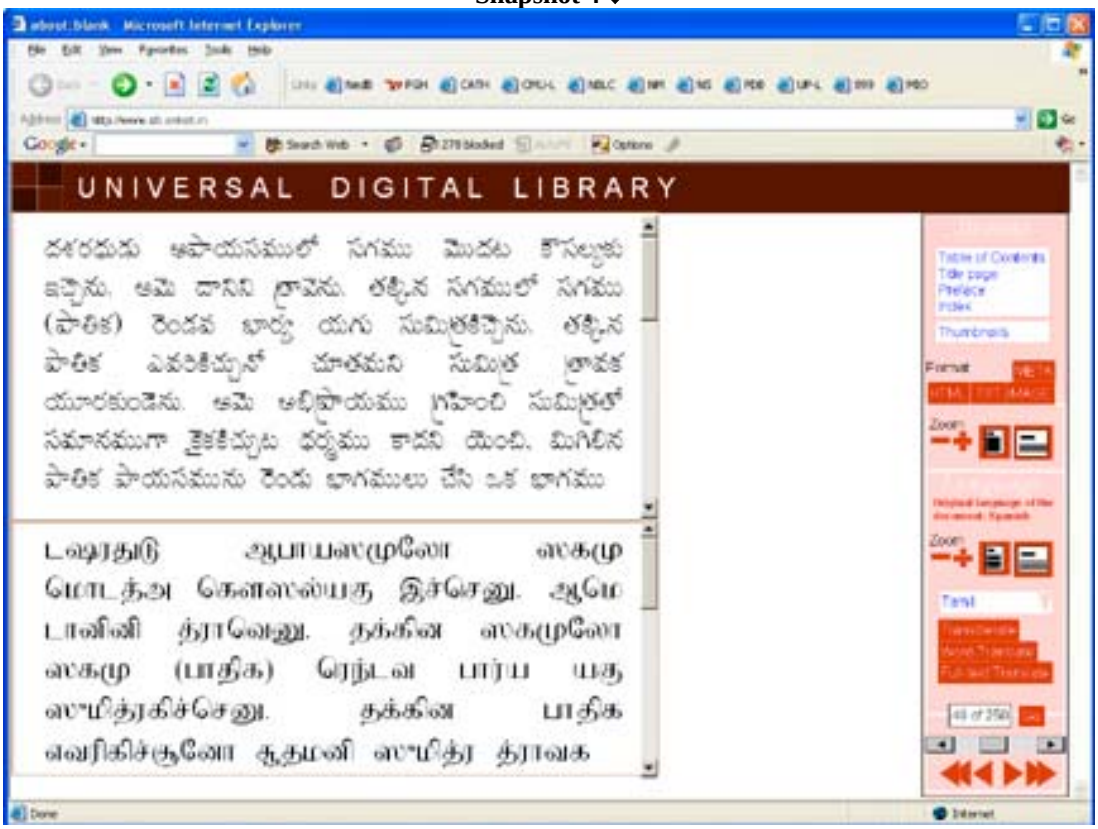
Snapshot 2 ↓



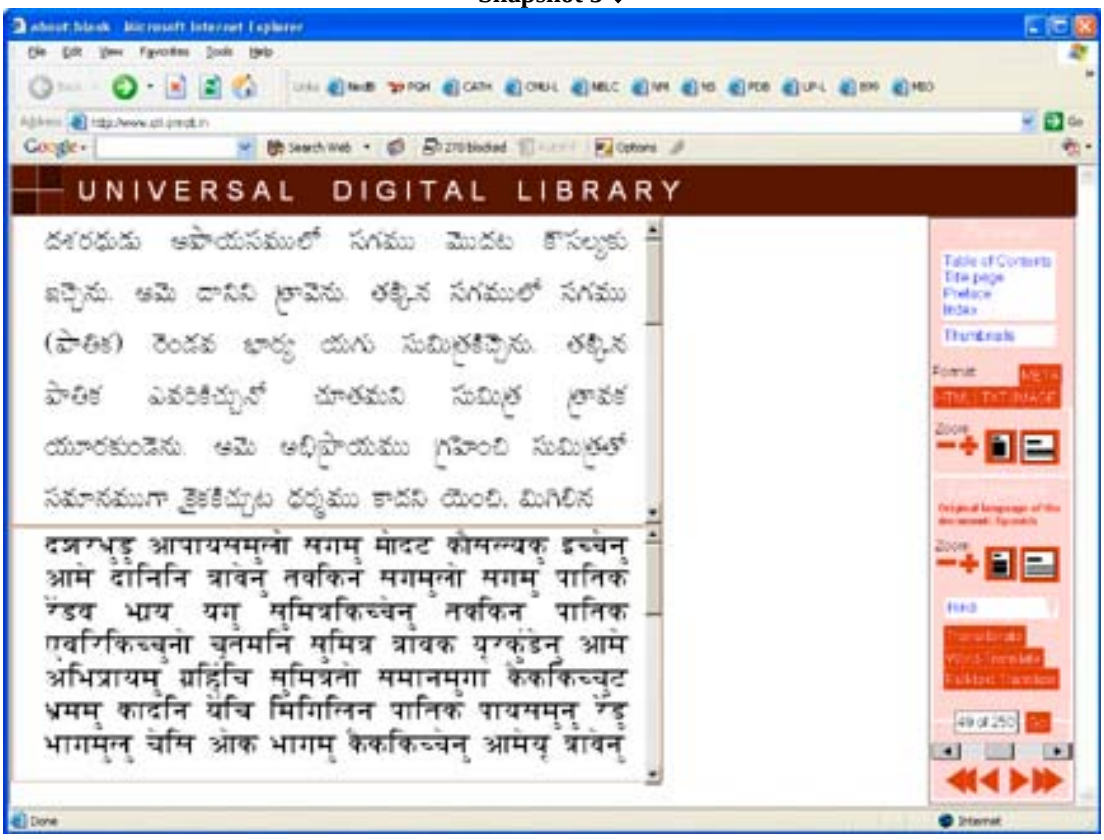
Snapshot 3 ↓



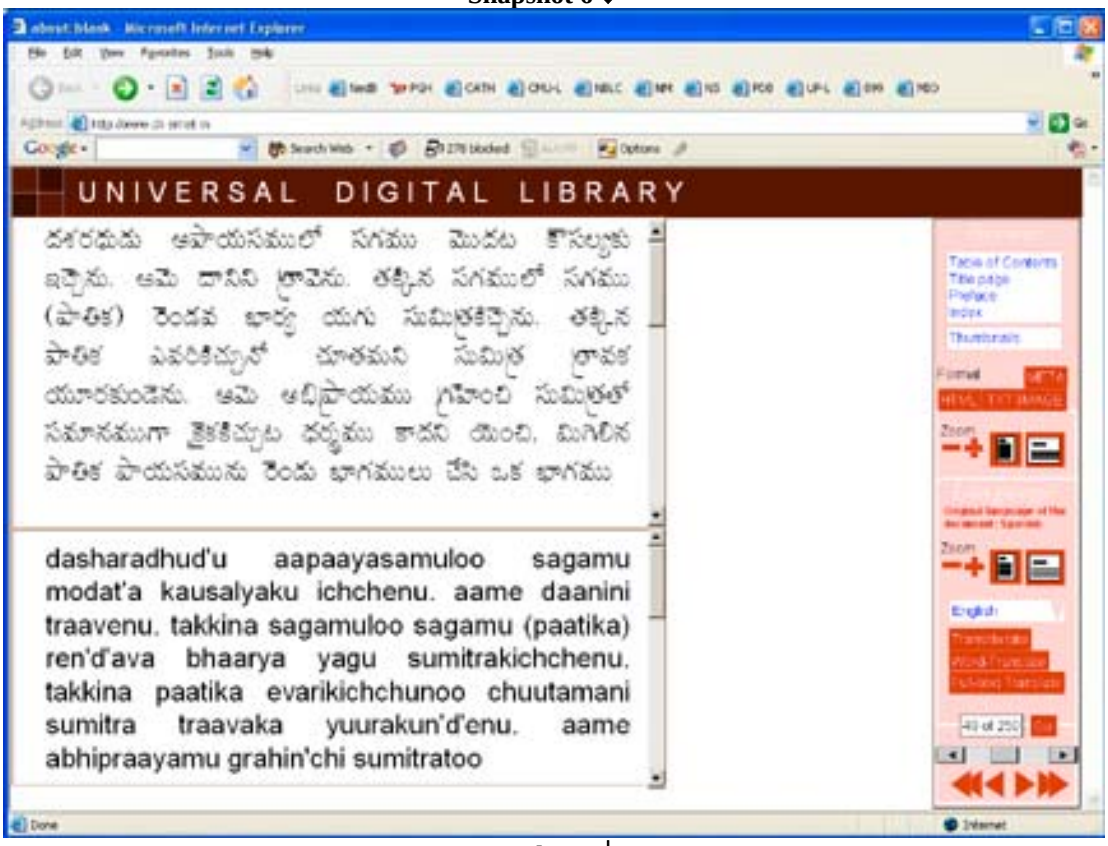
Snapshot 4 ↓



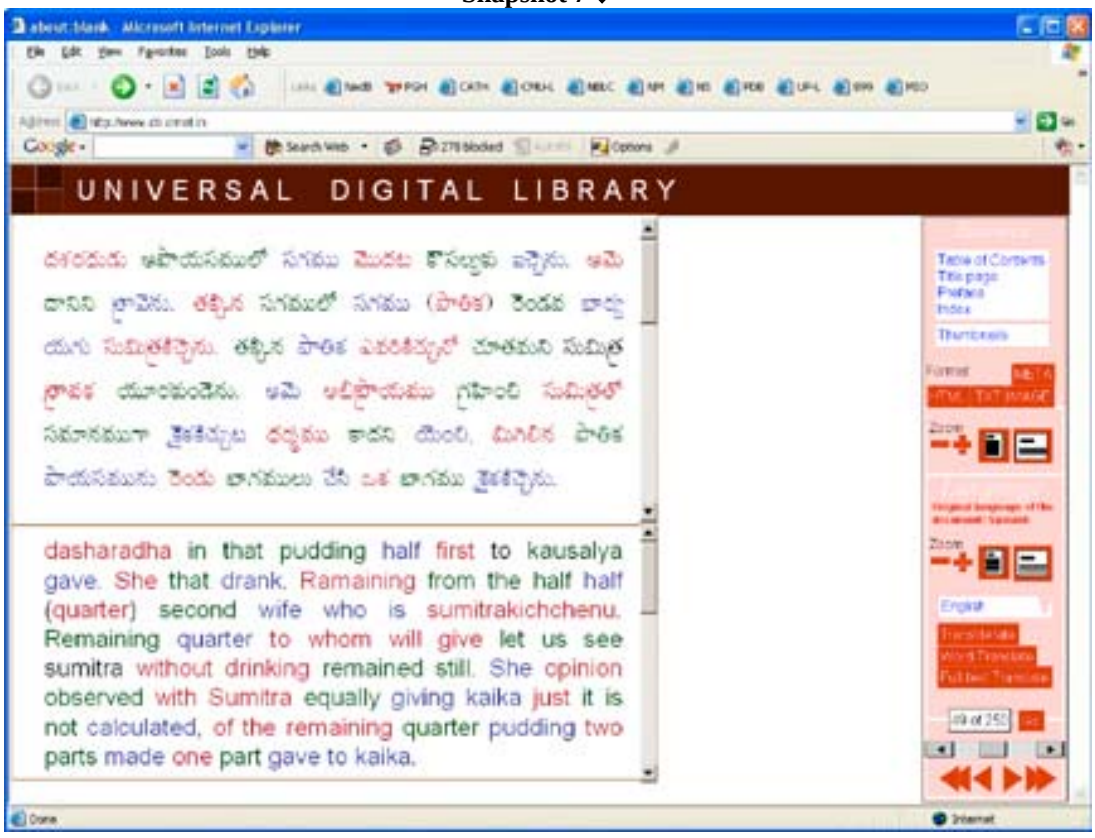
Snapshot 5 ↓



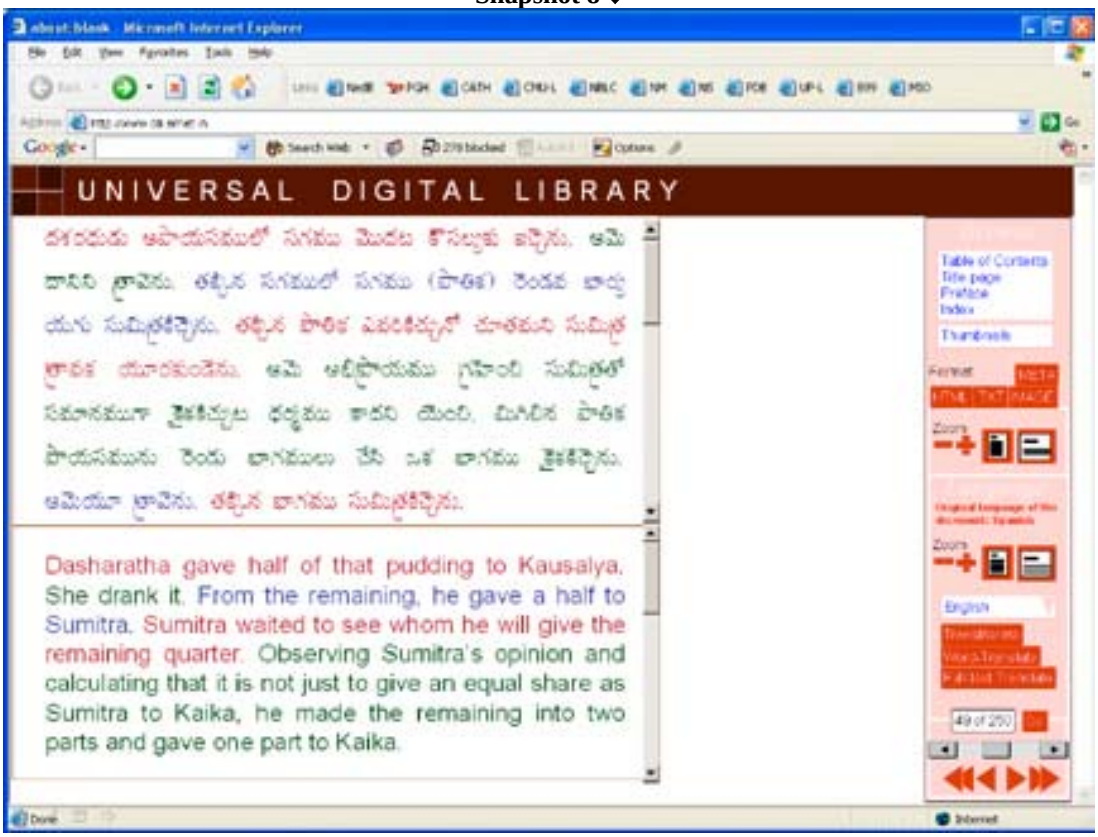
Snapshot 6 ↓



Snapshot 7 ↓



Snapshot 8 ↓



Conclusions

The multi-lingual book reader described in this paper is a simple to use and inexpensive tool and exploits the similarity between Indian languages. It is very useful for users who can understand their mother tongue or other Indian languages, but cannot read the script, and is good enough for an average reader who has the domain expertise. It can also be extended to multilingual search wherein either the documents or the queries can be translated in the same manner as is done for the reader.

Links to the DLI website

Digital Library of India: http://www.dli.ernet.in/
Example based machine translation for Indian languages: http://bharani.dli.ernet.in/ebmt/
Om transliteration scheme and integrated editor (Windows/Java standalone version for Linux, etc) and the web-interface: http://swati.dli.ernet.in/om/
Other DLI related publications may be found the website of the author NB: http://swati.dli.ernet.in/balki/

Acknowledgements

This work was supported by a grant from Microsoft India under Project Bhasha. The authors would like to acknowledge the support received from Microsoft and also the keen interest taken by Raveesh Gupta.

References

Balakrishnan, N. and R. Reddy (2004). Universal Digital Library - A test bed for Indian Language Technology Research. International Symposium on Machine Translation NLP and TSS (ISTRANS-2004), New Delhi, India.

Balakrishnan, N., R. Reddy, M. Ganapathiraju, et al. (2005). "Digital Library of India: A testbed for Indian Language Research." IEEE Technical Committee on Digital Libraries Bulletin: Special Issue on Asian Digital Library Research In press.

Balakrishnan, N., R. Reddy, M. Ganapathiraju, et al. (2004). Million Books to Web: Technological Challenges and Research Issues. Proc. Tamil Internet conference, Singapore.

Ganapathiraju, M., M. Balakrishnan, N. Balakrishnan, et al. (2005). OM: "One Tool for Many (Indian) Languages". ICUDL: International Conference on Universal Digital Library, Hangzhou.

Jayaraman, A., S. Sangani, M. Ganapathiraju, et al. (2004). OmSE: Tamil Search Engine. Proc. Tamil Internet conference, Singapore.

Porter, M. F. (1980). "A description of the most widely used English stemmer, An algorithm for suffix stripping." Program 14(3): 130-137.