OM: "One Tool for Many (Indian) Languages"

Madhavi Ganapathiraju[1], Mini Balakrishnan[2], N. Balakrishnan [2] and Raj Reddy[1]

Supercomputer Education and Research Centre, Indian Institute of Science, India
Language Technologies Institute, Carnegie Mellon University, USA
madhavi+@cs.cmu.edu, dli@dli.ernet.in, balki@dli.ernet.in, rr+@cmu.edu

Abstract

A large number of different languages are spoken in India, each language being the mother tongue of tens of millions of people. While the languages and scripts are distinct from each other, the grammar and the alphabet are similar to a large extent. One common feature is that all the Indian languages are phonetic in nature. In this paper we describe the development of a transliteration scheme *Om* which exploits this phonetic nature of the alphabet. Om uses ASCII characters to represent Indian language alphabets, and thus can be read directly in English, by a large number of users who cannot read script in other Indian Languages than their mother tongue. It is also useful in computer applications where local language tools are not yet available, such as email and chat. Another significant contribution presented in this paper is the development of a text editor for Indian languages that integrates the Om input for many Indian languages into a word processor such as Microsoft Winword®. The text editor is also developed on Java® platform that can run on UNIX machines as well. We propose this transliteration scheme as a possible standard for Indian language transliteration and keyboard entry.

Availability: http://swati.dli.ernet.in/om/, http://www.cs.cmu.edu/~madhavi/Om/

Contact: madhavi+@cs.cmu.edu, balki@dli.ernet.in, rr@cmu.edu

1. Inter-language transliteration is required for Indian languages

India is a nation with pluralistic culture, with 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. 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 familiar only with 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).

Language Technologies and PC Penetration in India

India is fast becoming a software superpower—the nation has over 3000 computer training institutes; software exports were about 6 billion US dollars in 2003, and are expected to grow to US$ 50B, which is 33% of total exports, very soon. However, net-surfers that were at 0.2% of the total population are expected to grow only up to 7% by 2006. The PC penetration rate is merely 1.4%. 68 million homes out of 408 million homes (17%) in the country have a TV, while only 22 million (5%) have a telephone; which is still much larger compared to the 1.4% penetration of a computer. Two most important influencing factors for this low computer usage by non software-professionals may be low income and illiteracy. The low income population in the country, which is a third of the total population, prefers to buy a television set rather than a PC because of the entertainment value, ease of use and the current non-utility of a PC in their everyday life.

At the time of the birth of independent India, about half a century ago, the Indian middle class was an insignificant minority; but the middle class is upwardly mobile. With the economic reforms brought about in the early 1990's, Indian middle class is growing at a rapid rate and is expected to reach 50% within a generation, and the poverty is expected to diminish to 15%. Complementing the economic growth rate, the new Indian middle class is filled with entrepreneurs who are spreading the power of information technology to the rural areas. Although the PC has not yet penetrated into rural homes, there are countless Internet facilities (called *cyber-cafés*) that are expanding similar in scope and impact to the public telephone booths in the rural areas. Low-end computers, costing about $100 to $200 are coming to the market (Simputer, Mobilis, Nova NetPC). Thus, irrespective of economic status, the power of information technology is expected to be available for Indian population very soon.

The second limiting factor in PC usage however, is non-availability of the operational software in native language, and the language barriers between people. While the development of an operating system in the native language is a solution, this is likely to be limited to only a couple of languages; further the development of natural language processing technologies would have to wait until the standardization of the digital representation; the porting of available scientific knowledge in the areas of natural language processing would face the bottleneck of a local expert in the native language. If the Indian language texts are instead available in parsable English-like texts, they would seem attractive to the international research community in language processing. Isolated development of digital representations for the different Indian languages may further widen the language barrier in the country.

Thus there is 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 summarization and statistical language modeling.

The representation must exploit the common alphabet of the various Indian languages. It must cater to the increasingly large number of people that can speak, but not read the native language—these people often can read another Indian language or English.

2. Prior transliteration methods built around standard keyboard

ITRANS is a representation of Indian language alphabet in terms of ASCII. (http://www.aczoom.com/itrans/). There are typically about 13–18 vowels and 36-54 consonants in Indian language—while there are only 26 letters in the English alphabet. Since Indian text is composed of syllabic units rather than individual alphabetic letters, ITRANS uses combinations of two or more letters of English alphabet to represent an Indian language syllable. However, there being multiple sounds in Indian languages corresponding to the same English letter, not all Indian syllables can be represented by logical combinations of English alphabet. Hence, ITRANS uses non-alphabetic characters such as "[", "\", "'" in some of the syllables. These combinations are not logical and are not easy to remember and recall. ITRANS notation is also case dependent—it uses capital and small letters to represent different language syllables. Another major drawback of ITRANS proposed so far is that the same Indian Language Character

can be represented in more than one way using lower and uppercase letters, making the transliterated non-uniform across people.

Unicode standardization captures the commonality in the alphabet of various Indian languages, but it does provide for an input mechanism. It does not provide for a logical mechanism of applying the Language parsing algorithms on texts encoded in this format. The lexical ordering of the Indian languages cannot be applied in a logical fashion. The representation does not automatically transliterate to English, which is an important requirement as discussed above.

A very significant contribution in this area is that of the Acharya group at that Indian Institute of Technology, Madras (http://acharya.iitm.ac.in/). They have developed a representation that preserves the syllabic and phonetic nature of Indian languages and also preserves lexical ordering. However, the representation is only machine-readable, but the input and English transliteration are still based on ITRANS. This is very good for internal representation and also for lexical ordering and syllabic parsing such as finding palindromes in the text. But absence of a mapping to ASCII makes it in-adaptable to standard Language parsing applications.

To overcome the drawbacks of ITRANS we have redesigned a novel mapping scheme called Om, which is no longer a transliteration mechanism alone, but a platform over which many other Indian language applications have been built; the details of which are described in the rest of this paper.

3. Om transliteration: unified representation for Indian languages

Om uses the same representation both for keyboard input and formation and representation. It is similar to ITRANS in that it uses combinations of English alphabet to represent Indians syllables. However, it is case independent, and avoids excessive use of non alphabetic characters; where used they are consistent. Further, the English alphabet combinations are designed such that they are easy to remember at the time of input with standard keyboard and also natural to read like English. The case independent representation allows use of sentence and title case writing in a natural fashion; further, the texts are highly readable than their ITRANS counterparts. It may be seen from any ITRANS text that the large mixture of capital and small letters, and not an alphabetic characters leave it highly difficult to read.

Om's features enhance the usability and readability, it has been designed on the following principles: (i) easy readability (ii) case-insensitive mapping: while preserving readability, this

feature allows the use of standard natural language processing tools for parsing and information retrieval to be directly applied to the Indian language Texts and (iii) phonetic mapping, as much as possible: this makes it easier for the user to remember the key combinations for different Indian characters ASCII representation may be used simply as a means of typing the text with standard keyboard. (iv) Om separates the storage that is in ASCII and the rendering that is dependent on the fonts chosen. This paves the way for a language independent universal representation; a fact that had been exploited in multilingual search engines. 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 or Acharya, 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. India being a multi-lingual country, and inter-mixed population, often the people can speak and understand more than one Indian language and also English. Hence even in the absence of Om to native font converters, people around the globe can type and publish texts in Om scheme which can be read and understood by many even when they cannot read native script. The readability criterion that is benefited from the case-insensitive phonetic mapping thus proves very useful.

The Om mapping tables for many Indian languages are shown at http://swati.dli.ernet.in/Om/. The table also shows the mapping for the characters, and some sample Om texts. Mapping table for Kannada is shown below as an example.

A fully-filled mapping table with Om characters as columns and different Indian languages as rows is also created: in this table, whereever a character present in one Indian language alphabet is not present in the alphabet of another, it is substituted with a similar sounding character in the latter language. For example, the Tamil character *n*- is not present in Telugu, and hence the Om character *n*- is substituted with *n'* in Telugu.

Kannada Character Mapping

ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ	ಎ	ಏ	ಐ	ಒ	ಓ	ಔ	ಅಂ	ಅಃ
a	aa	i	ii	u	uu	rx	rx-	e	ei	ai	o	oo	au	n'	:

ಕೃ	ಖ	ಗ	ಘ	ಙ
ಚ	ಛ	ಜ	ರ್ಾ ಽ	ಞ
ಟ	ಠ	ಡ	ಢ	ಣ
ತ	ಥ	ದ	ಧ	ನ
ೈ	ಫ	ಬ	ಭ	ಮ

ka	kha	ga	gha	ng-a
cha	chha	ja	jha	nj-a
t'a	t'ha	d'a	d'ha	nd-a
ta	tha	da	dha	na
pa	pha	ba	bha	ma

ಯ	ರ	ಲ	ವ	ಶ	ಷ	ಸ	ಹ	ಳ	ಕ್ಷ
ya	ra	la	va	sha	shha	sa	ha	l'a	qs-a

4. Word processor

An integrated transliteration package that accepts Om ASCII keystrokes as input and maps them to native fonts has been developed. The script in any one of the chosen true type fonts is sent to MS word 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 also be saved in plain ASCII and Unicode formats. The tool also integrates with email clients on the windows platform. A web-interface with similar functionality has also been developed. The text may be saved as Om text, native font text or in Unicode. This does not support formatting explicitly but can be independently opened in MS Word like applications for such functionality.

Easy support for new languages

A mapping table between Om symbols and the grlyphs of the font of the new language is required. Once this is provided, it is only a matter of a few minutes to integrate this new language into the package. All the other features of transliteration to other languages and use of word-editing features of Microsoft word are avialble after the integration of the new font into the package. Currently, the Om transliteration package supports eight Indian languages:

Key in the input as we speak

The most notable feature of the OM transliteration package is we can key in the input data just the way it sounds when we speak. For example if we have to key in ‘Bharat’ just type ‘bhaarat’.

Uses lowercase English alphabets and some special characters

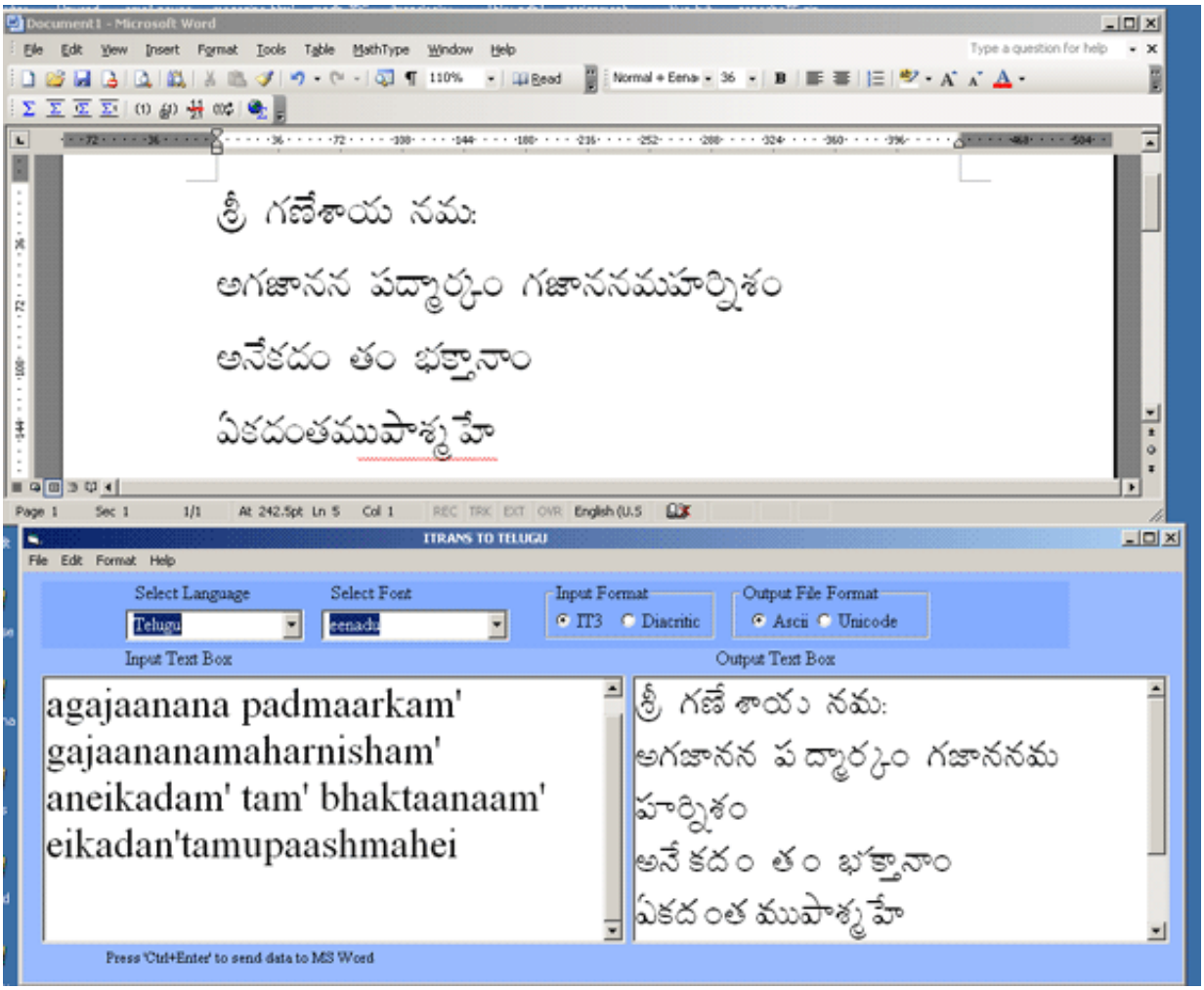
The use of lowercase letters provides awesome power to adapt language modeling tools such as stemmer, translation etc. The special characters used in OM are ‘, *, ~,

Switch between the languages at the click of a mouse

The option to choose any language and font is incorporated in the interface of OM by which switching from one language or font to the other is made easy.

Saves the output in ASCII and Unicode format

The file menu of the interface provides an option to save the input as well as the output, so that the user can import it later for future use.



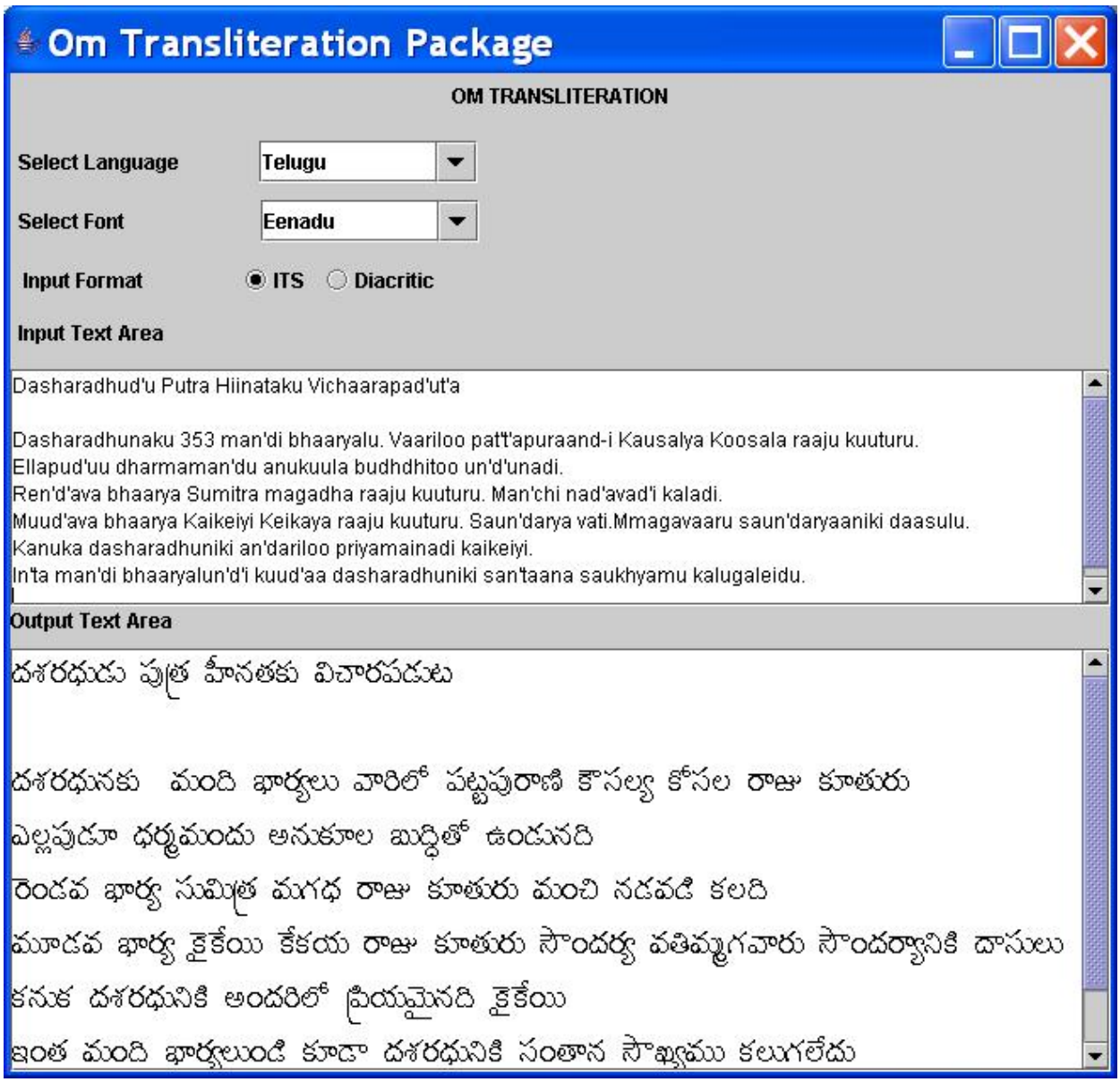
Exchange email in Indian languages

This feature lets the user to send electronic mail in plain text in Indian languages

Integration with Microsoft® Winword (MSWord)

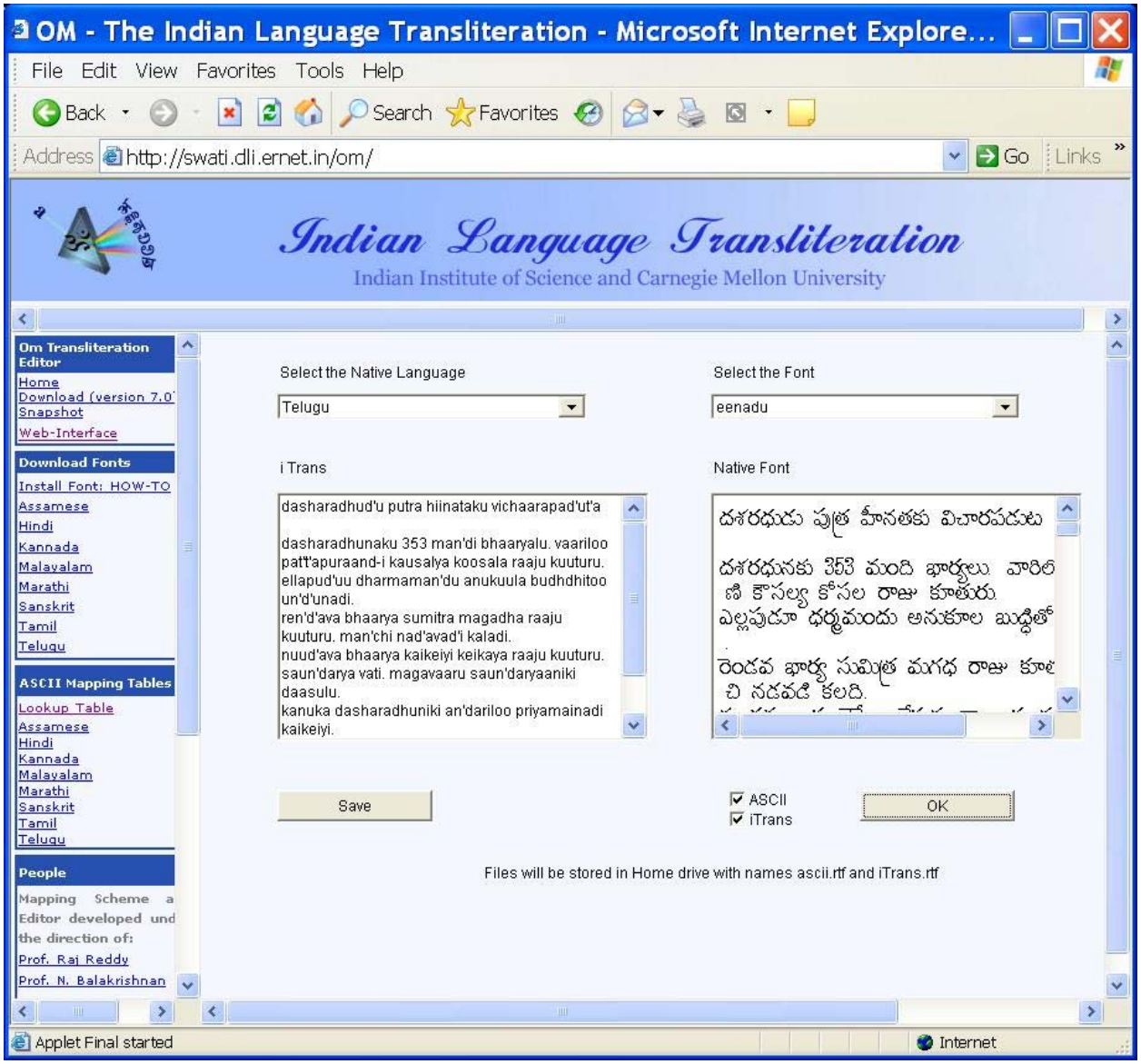
The output can be exported to MSWord allowing users to take advantage of all the features in MSWord provided this application is present in the user's computer.

Platform independent package in Java



Web Interface

For those who wish to create content using a web interface, without the need to install the package locally, a java based web interface is also available. (http://swati.dli.ernet.in/om/ and http://www.cs.cmu.edu/madhavi/Om). The web interfce creates the output in plain text format, which may be opened in MSWord with the appropriate font selection, thereby using the full functionality of MSWord for the Indian language text editing.



5. Conclusions and Future Work

A transliteration and keyboard entry scheme for Indian languages called *Om* has been described in this paper. Integrated text editing tools, for both Windows and Linux platforms, and also a web service for the same, have been presented. The editor allows entry of text using Om mapping scheme using a standard keyboard, and converts the text to native language fonts. The editor, and the design of Om, also allow transliteration of the text from one Indian language to another. All the tools are available freely for use, download and hosting. Supplementary material consisting of all the mapping tables and inter-conversions between different languages is available on the website.

6. Availability: Free for download and hosting

The Om transliteration mapping and integrated editor are available for download at http://swati.dli.ernet.in/om/ and http://www.cs.cmu.edu/~madhavi/Om. The tools have been used extensively for data entry for texts that feed into applications such as machine translation and optical character recognition. It has also been used purely for content creation by outside community. An example may be seen at the magazine section of www.telugumn.org, where the story of Ramayanam has been created using this software. The integrated editor will also be provided for hosting at any website free of cost or use, such as done at www.telugumn.org. The integrated editor is available for windows and linux platforms.

7. National Standard:

In India, the Ministry of Communication and Information Technology under its “Technology Development for Indian Languages” (TDIL) has been working on evolving national standards for representation and localization. http://tdil.mit.gov.in/homepage.asp describes some of the past and present attempts in standardization including the use and issues connected with the ISCII, UNICODE, INSFOC and INSROT. There have been many scattered attempts in this direction by some of the academic institutions and research organizations across the country. It is proposed to have a national conference of all the language and computer experts to brain storm and decide on evolving an acceptable national standard like Om so that in all our future endeavors language as a barrier to ICT applications reaching the Indian rural populations and to the success of our E-Governance exercises would removed.

8. Acknowledgments

Om transliteration and integrated editor have been developed by a large number of people at the Multimedia Systems Lab at the Supercomputer Education and Research Centre, Indian Institute of Science and at the ISRI, Carnegie Mellon University. Of particular mention are the names of Sravan Kumar, Jiju Verghese, Sheik, Tina Joseph and Umi who contributed towards specific Indian languages. Jiju Verghese developed the web interface and the Java standalone version has been developed by Atul Kumar.

9. References

The statistics in the first 2 sections are collected from the following pages:

1. http://www.forachange.co.uk/index.php?stoid=168
2. http://www.davidhwells.com/PhotoEssays/globalindia/middleclas/default.html
3. http://www.etstrategicmarketing.com/smJune-July2/forum.htm
4. http://www.ogilvy.com/viewpoint/view_ko.php?id=16216&iMagaId=6
5. http://www.siliconindia.com/shownewsdata.asp?newsno=28077&newscat=Technology
6. http://www.engr.mun.ca/~adluri/telugu/language/usage/grierson/introduction.html