Cover Story

Hareesh N Nampoothiri
University of Kerala, Thiruvananthapuram

Desi Language Computing - on the Rise

English was the first language that got placed in modern computer systems and naturally got accommodated exclusively, to the disadvantage of the other world languages. From the mnemonics used in assembly language, to the programming language keywords, to operating system commands, English embedded itself. Some early programming languages like COBOL almost sounded like English of nonnative speakers of the language. It is easy to weave an Anglo-centric conspiracy story, but in all fairness to the professionals of the yesteryears, it must be remembered that computers were not foreseen then as gizmo gadgets that ordinary citizens all over the world would own. As the popularity of the notebooks, netbooks, and mobile devices shot up, the language problem began to take a central stage and naturally multiple solutions began to emerge. Perhaps the turning point in language computing is the emergence of the Unicode. Unicode is simply a computing industry standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems[1]. It set the stage for an organized development of a large number of linguistic computing issues.

Even though the first version of Unicode was introduced in October 1991, it became popular only in the last decade. As of now, Unicode supports a long list of languages including Indian languages such as Bengali, Hindi, Kannada, Malayalam, Oriya, Tamil, Telugu etc. Now software developers come up with different language packs for different regions and computers are becoming truly *desi* in this aspect. An example is Microsoft's CLIP (Caption Language Interface Pack) for Visual Studio 2010 in which the author was also associated for developing a language interface pack.

Apart from reaching a wider audience through incorporating as many languages as possible, Unicode also opens a wide range of possibilities for developers and service providers to come up with language-based tools and applications for common man. It is not surprising that Google is the one in the lead, tapping the possibilities in this sector. We introduce below a few of the language-based tools from Google.

Google Translate

Google Translate is a free translation service from Google, which provides instant translations between 65 different languages (as of Apr 2012) including some of the major Indian languages like Bengali, Gujarati, Hindi, Tamil, Telugu, and Urdu. Google Translation enables the users to translate words, paragraphs of text, or a whole website (using the Translator toolkit) from one language to another. According to Google the service aims to make information universally accessible and useful, regardless of the language in which it's written[3].

How does it work?

Google describes the working of Google Translate as follows: *When Google Translate generates a translation, it looks for patterns in hundreds of millions of documents to help decide on the best translation for you. By detecting patterns in documents that have already been translated by human translators, Google Translate can make intelligent guesses as to what an appropriate translation should be. This process of seeking patterns in large amounts of text is called "statistical machine translation". Since the translations are generated by machines, not all translation will be perfect. The more human-translated documents that Google Translate can analise in a specific language, the better the translation quality will be. This is why translation accuracy will sometimes vary across languages*[3].

In Practice

Let's see how it becomes useful in practice by trying to translate a simple paragraph from English to Hindi (Fig. 1). Of course, it

What is Unicode?

In early days, there were many different encoding systems for characters used in computers. These encoding systems used to conflict with one another. That is, two encoding systems may use the same number to represent two different characters or they may use different numbers for the same character. As a result, any given computer was required to support many different encoding systems and even after that the chances of getting data corrupted was very high.

To solve this issue, Unicode provides a unique number for every character irrespective of the platform, application, or language. The Unicode Standard has been adopted by most of the leading players of the industry such as Apple, Microsoft, Oracle, IBM, Sun etc. Also it is required by modern standards such as XML, Java, JavaScript, WML etc. It is supported in many operating systems (including Linux distributions), all modern browsers, most of the recent versions of office suites, and many other applications.

The Unicode Consortium, a non-profit organization, is dedicated to develop, extend, and promote use of the Unicode standard. According to them the advantage of using Unicode is:

Incorporating Unicode into client-server or multi-tiered applications and websites offers significant cost savings over the use of legacy character sets. Unicode enables a single software product or a single website to be targeted across multiple platforms, languages and countries without re-engineering. It allows data to be transported through many different systems without corruption[2].

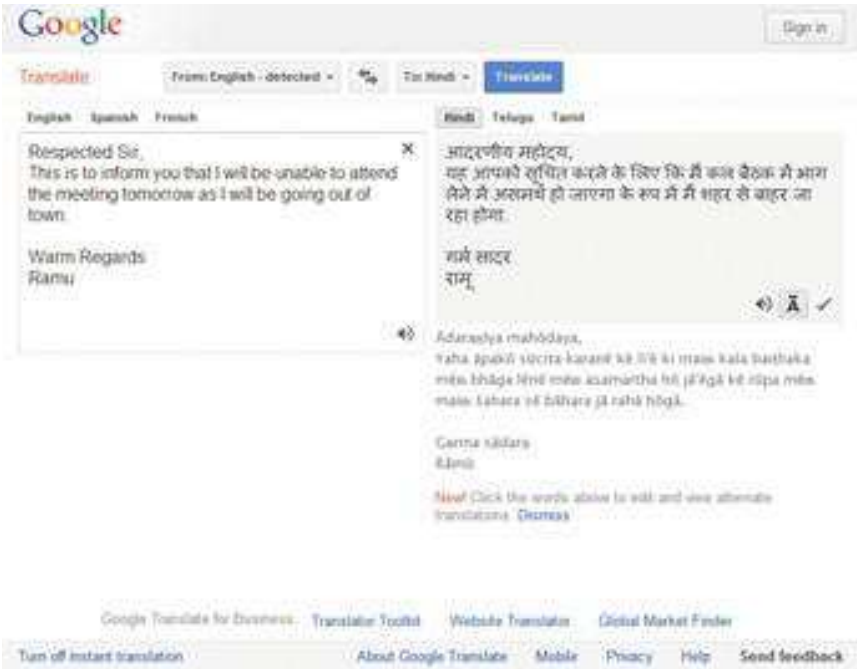

Fig. 1: Google Translator Page: http://translate.google.com/

does not produce a grammatically correct translation, but it does produce a useful text in Hindi. Apart from providing the translation of the text, it also provides the phonetic rendition of the text in English. One can hear the translated text by clicking the speaker icon.

There is also an option to rate the resulting translation by clicking the tick mark. One can rate a particular translation as Helpful, Not helpful, or Offensive. The tool also offers alternative translations and an option to re-order blocks of words for reconstructing the translated sentence (Fig. 2).

So what about translating from one Indian language to English? For that we need to type-in the text in the required Indian language. There is another tool from Google, Google Transliteration (still in labs) that will help you to type in other languages without learning the actual keys corresponding to the alphabets of that particular language. Here we will type 'mera bhArath mahaan' to get 'मेरा भारत महान' in Hindi.

Fig. 2: The tool suggests alternative translation when the user click and hold on a block of words

The transliteration window (Fig. 3) provides required options to edit and format the text. Google provides transliteration API that helps the developers to enable transliteration facilities in their websites. The transliteration API is incorporated in Google Translate as well. When a language other than English is selected in Google Translate source window, an option to enable phonetic typing will be available. By enabling the option, one can directly type-in the required text in the Translate source window itself. Another option is to copy-paste the typed text from Google Transliteration window.

Note: Apart from Google Transliteration, there are many online and offline tools available, that will help you to type-in text in Indian languages. For Windows-based systems one may use Indic Input 2 (for Windows Vista / 7) or Indic Input 1 (for Windows XP). By installing this tool, one can type-in text in any text editor (such as Notepad, Wordpad, LibreOffice, Writer etc.) by enabling the phonetic keyboard and selecting the appropriate Unicode font. The tool can be downloaded freely from the BhashaIndia website. URL: http://bhashaindia.com/Downloads/

Here are some amusing translation examples – the lyrics of a Hindi film song (Fig. 4) and our national anthem (Fig. 5). When the Hindi film song lyrics are translated, the tool produces acceptable results but the translation for the national anthem is amusing, to say the least. In short, for simple functional sentences it produces better translations and for creative writings (such as poems) the results may not be of utility.

Developers can integrate the application in the websites and it

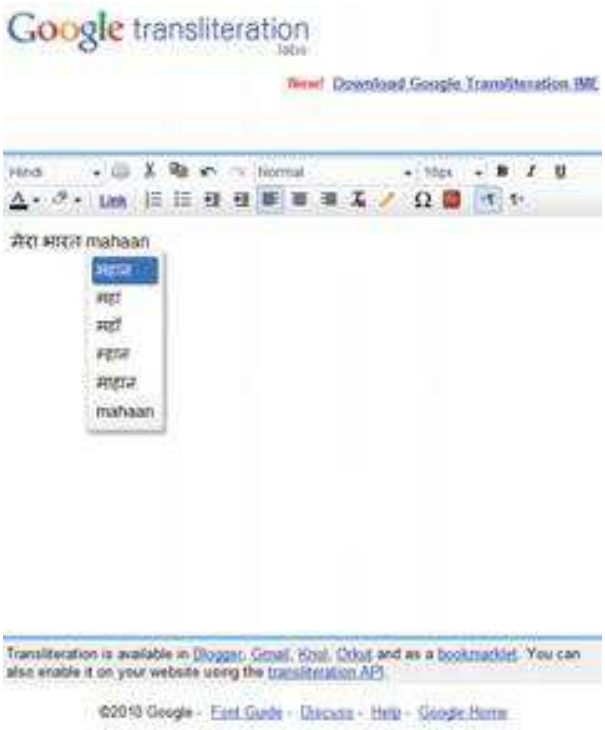

Fig. 3: Google Transliteration window

Fig. 4: Hindi film song lyrics translated to English

automatically translates the website to another language according to the choice selected by the user (Fig. 6). Even though the tool does not produce acceptable results all the time, it will be useful in translating websites to local languages (or foreign languages) using the Translator Toolkit provided by Google. At least the users will get some idea about the contents of the website instead of seeing the website in some alien language.

What if the website does not provide a translation option by default? Still, it is possible to view the website in a language of your choice. For example, Computer Society of India website does not have an option to switch between languages. But still it is possible to display the website in Hindi or in any one of the 65 languages provided by Google Translator. If you wish to see the CSI website in Hindi, enter the following URL in the address bar: http://translate.google.com/translate?tl=hi&u=http://www.csi-india.org

The `tl` (target language) parameter corresponds to the language of your choice (`hi` for Hindi, `tl` for Tamil, `bn` for Bengali and so on) and `u` is the URL of the website you wish to translate. The translated version of the CSI website is shown in in Fig. 7.

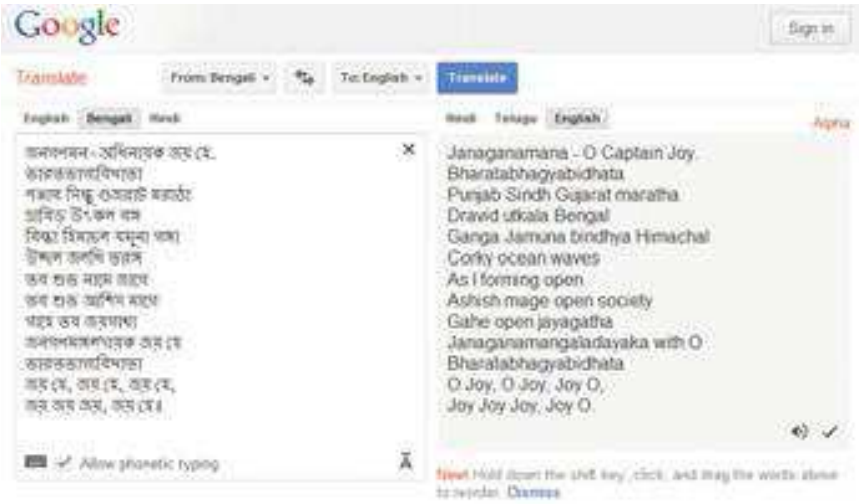

Fig. 5: National Anthem translated from Bengali to English

Fig. 6: Sample website with Google Translate enabled using the API.
When the user scrolls over the text, the original text will be displayed as a tool-tip dialogue

Alternatively, you may install Google Toolbar or get a bookmark for your language from the Tools and Resources page. URL: http://translate.google.com/translate_tools

Mobiles & Tablets Too Go *Desi!*

It is not happening with computers alone. Most of the modern mobile devices (Smartphones, tablets etc.) boast the power of computers we had three decades back. Apple Lisa[4] (released in Jan 1983), the first personal computer which offered GUI, had the processing power of Motorola 68000 @ 5 MHz. Now the medium range smartphone, Motorola Defy has 800 Mhz processor. If the memory of Apple Lisa was 1 MB RAM (In Lisa 2 only Apple introduced 10MB internal hard disk drive!), Motorola Defy has 512 MB RAM, 2 GB internal storage, and it supports microSDHC upto 32 GB! The tablets currently available in the market are even more powerful and we may consider them as minicomputers, only difference being the lack of input devices like keyboard and mouse (Of course, they permit to add them too via Bluetooth or USB!). Mobile devices are becoming more popular and the manufacturers are trying to reach mass public by incorporating local language support in their mobile devices. Clearly, the *'desification'* is not going to happen in computers alone but it will extend to mobile devices as well.

Many of the devices produced by various cell phone/tablet manufacturers like Nokia, Sony, Samsung, LG, Motorola etc. already allow the users to select a language for the phone interface. Entering and displaying Indic languages directly in mobile devices (for sending messages, for contact details, for writing notes etc.) is still in the development stages. Apple, the leading mobile device manufacturer, provides local language support in

Fig. 7: CSI website translated to Hindi

Fig. 8: Hindi text displayed on an Android mobile phone

their iPhones and iPads based on iOS mobile operating system. Even though many of the other devices from various manufacturers do not have native support for Unicode, there are device specific work-arounds available for incorporating Unicode functionality in those mobile devices, especially for devices based on Android platform. Android-based devices from Samsung, LG etc. comes with support for Indian languages by default. In some mobiles, in the keypad itself, the Hindi alphabets are printed along with English alphabets to make entering the text easy as possible. Fig. 8 shows a low-end Android mobile phone from LG using Google Translate. The text produced is then copy-pasted to a message and send. If the party receiving the message has a mobile device with Unicode support, then the text will be rendered correctly or else the receiver will get a series of squares instead of the actual message.

It is very obvious that developments in Indian language computing have moved very much to web and mobile platform rather than as stand-alone applications on PCs. The demand for these tools now arise from the common man and not from business or universities. That explains the vibrancy of this field in this current times.

References

[1] Wikipedia – Unicode
http://en.wikipedia.org/wiki/Unicode
[2] What is Unicode?
http://www.unicode.org/standard/WhatIsUnicode.html
[3] About Google Translate
http://translate.google.com/about/intl/en_ALL/
[4] Wikipedia - Apple Lisa
http://en.wikipedia.org/wiki/Apple_Lisa

About the Author

Hareesh N Nampoothiri is a visual design consultant with an experience of more than a decade and worked with government organizations like C-DIT, C-DAC, University of Kerala, and other private organizations. Currently, he is doing interdisciplinary research in ethnic elements in visual design in computer media. He is an author of two books on graphic design and a regular contributor in leading technology magazines including CSI Communications. Kathakli, blogging, and photography are his passions. He has directed a documentary feature on Kathakali and also directed an educational video production for IGNOU, New Delhi.