Intelligent Text Input and Superior Display on Information Appliances: Asia-Pacific Region

Sara Hirschhorn
Vice President, Sales and Marketing, SlangSoft
S.Senthil Nathan
Linguistic Engineer, SlangSoft

ABSTRCT

Internet Appliances such as mobile phones and set-top boxes are increasingly becoming a standard means by which to access the Net, especially in Asia where PC penetration is notoriously low and communications infrastructure poor. Internet Appliances are routinely less expensive than PCs, presenting a more cost effective and reliable way to get online.

In order to realize the tremendous growth of Internet users we, as software developers, need to make inputting and displaying text in any language as fast and intuitive as possible, via any device. For us, the User Interface is the focal point of interaction between end-users and applications. Appliances and Applications that allow for easy and stress-free input and display - which are after all, the key interactions involved with UI -- will unquestionably have a competitive edge over those that don't. A complaint often voiced by consumers is that entering text on small digit keypads is annoying and frustrating.

The popularity of Net appliances necessitates that all text input and display applications targeting Asian markets posses a small footprint, so as not to compromise the performance or speed of small enabled devices that are memory constrained. The cross-platform usage is also important.

For input of Indian languages, the process is simpler but still involves a variety of input methods. Even for the PC, there are number of non-standard keyboard layouts and encoding. With the exception of Tamil, no standard has been developed by Governmental or other bodies to govern input of other Indic languages. The inscrypt layouts of Department of Electronics, India, used in iLeap and some other software, are considered the de facto standards for pan-Indic software developers but for other Net Appliances, no informal or formal standard exists.

Most of the technology that is available today to enable fast input of text from Internet employs linguistic dictionaries. However, since maintaining a super small footprint is critical to any embedded device and most Net appliances, efforts are constantly made to increase efficiency of fast text input by offering more functionality with smaller RAM requirements. Slangsoft's iTID Platform, for example, includes linguistic data with frequency of usage which relies on a proprietary algorithm which combines root and word inflection rules to greatly expand extremely compact wordlists for each language to keep the size of the linguistic data very small.

The future holds the promise that computing capabilities will be contained in everything that human beings interact with such as cars, furniture, and clothing. The manufacturers of these

items, along with the makers of applications that run on them, will be challenged to provide fast, intuitive ways for people to interact with them.

INTRODUCTION

With a population of over three billion people, the 23 countries of the Asia-Pacific region represent a rapidly growing and lucrative segment of the global Internet market. The number of Internet users was projected to have reached 48.7 billion by the end of 2000 and is expected to reach 173 billion by the end of 2004 (eMarketer).

Internet Appliances such as mobile phones and set-top boxes are increasingly becoming a standard means by which to access the Net, especially in Asia where PC penetration is notoriously low and communications infrastructure poor. Internet Appliances are routinely less expensive than PCs, presenting a more cost effective and reliable way to get online. China is an excellent example: it is poised for a boom in users accessing the Net via Internet Appliances not only because it represents the world's third largest wireless population, but also because televisions out number PC's 25 to 1, with 80 million cable subscribers (eMarketer). You can find a similarity between China and India as far as the usage of television is concerned.

The pattern of growth of Net and wireless usage in Asia varies. In Japan, accessing the Web from mobile phones has become so popular that mobile phone users are now the fastest growing segment of the online population; currently 10 million people go online through mobile connections (Japanese Ministry of Post and Telecommunications). In India, use of Net and wireless services is still a limited one but significant growth is evident. While Japan provides the benefits of mature market, India provides the benefits of emerging markets. In India, the larger markets lie in the Western and the Southern states. In South, Tamil Nadu is one of leading tech savvy state. In India, Tamil speakers represent a significant market for internet and communication tools. Also, in Singapore, Malaysia and the West, Tamil speakers are slowly adapting the newer infocomm technologies. In short, even among Indian languages, Tamil is one of the leading language for ifocomm media. This fact attracts multilingual technology holding companies like Slangsoft to the Tamil market.

FROM FORECAST TO FRUITION

In order to realize the tremendous growth of Internet users we, as software developers, need to make inputting and displaying text in any language as fast and intuitive as possible, via any device.

Our idea is predicated on the belief that the User Interface is the focal point of interaction between end-users and applications. Appliances and Applications that allow for easy and stress-free input and display - which are after all, the key interactions involved with UI -- will unquestionably have a competitive edge over those that don't. Although English is the dominant technical and business language in Asia, the vast majority of users feel most comfortable inputting and viewing text in their native language.

The process by which users enter text on these devices is of equal importance. A complaint often voiced by consumers is that entering text on small digit keypads is annoying and frustrating. The multiple keypresses necessary to enter text on small digit keypads are enough to make some people want to throw their mobile phone out the window. If you've ever tried sending a text message from a mobile phone that isn't equipped with some type of fast or smart input, you know what I'm talking about.

THE CHALLENGE OF ASIAN LANGUAGES

Despite the radical diversity of peoples and languages in Asia, there are some common trends that have implications for linguistic software no matter which Asian language is inputted and displayed. For example, the popularity of Net appliances necessitates that all text input and display applications targeting Asian markets posses a small footprint, so as not to compromise the performance or speed of small enabled devices that are memory constrained.

To keep in step with the trend toward ubiquitous computing and the wide variety of platforms currently in use, any input and display software targeted at Asian users must be cross platform. Any display software must provide for accurate visualization and layout of text, especially for representation of ideographic languages. A rendering engine must ensure adjustment of characters to device screen size, aspect ratio, pixel dimensions and color depth and should be able to ensure optimized font appearance on TV screens and LCD displays.

In addition to the complexities associated with the display of ideographic languages, there are serious intricacies related to the display of South East Asian Languages such as Thai and Indic languages. There are 10 major scripts used in India to represent countless spoken languages. While there are major distinctions between the 10 scripts, a shared characteristic that leads to difficulty in the display of these languages are ligatures. A ligature is the end result of the combination or merging of two characters into another character. The rules stipulating how and when characters combine can present a significant obstacle to the accurate display of the Indic scripts. This impediment can be overcome through the use of advanced algorithms.

BARE NECESSITIES

Before attempting to improve user experience we must understand what functions native speakers of Asian languages expect to encounter when they enter text. Just as an English speaker would find it strange and unnatural to only be able to use lower case letters when entering text, there are several utilities that Asian language speakers have come to count on, such as, the ability to enter English words in the midst of a sentence written in Tamil.

Many of the languages spoken in Asia, including Chinese, Japanese and Korean, are ideographic languages. Ideographic languages use a system of pictures, symbols or characters to represent an object or an idea, but not necessarily an exact word. A system of phonetic input is used to input the basic sounds that compose the different characters or ideographs. Chinese, Japanese and to a lesser degree, Korean have a phonetic script that is used to input ideographic characters. For CJK languages several keystrokes are entered before conversion is done replacing these keystrokes with text. When the user inputs text, it is displayed in a noncommitted form and the user has to correct and/or accept this text, the text to be converted appears underlined. Composition is the term used to define the process of entering text that has not been confirmed by the user to be correct. Text that has been confirmed by the user is said to be committed. The user initiates conversion by pressing the space bar or a designated conversion key. The conversion engine uses complex logic to convert text input into another form (ideographic characters). Conversion engines come in different formats, all of them will define the pronunciation for ideographic characters. To ensure the most accurate conversion possible, the conversion engine should employ additional information such as, part of speech, verb and adjective conjugation, and frequency of usage data. Once the user has initiated conversion, the text that has been phonetically inputted is displayed as ideographic characters. The text displayed is still "composition" because the user still needs to confirm that the text displayed is what was intended.

Converting phonetic input into ideographic characters can lead to frequent mistakes, making proofreading or committing to the displayed text, a major aspect of entering text. To aid the user in this process a list of all possible ideographic matches for the phonetic input should be provided. This is a function that a native speaker of a CJK language would expect to encounter when inputting text. It is referred to as a "homonym lookup window" or a "list of candidates". To confirm that the displayed ideographic characters represent the meaning the user intended, the user presses on the space bar or simply continues typing. Confirmed text is displayed without the previous underlining.

Another must-have is the choice between methods of input. For example, in Japanese there are two methods of inputting text. One way is to enter text using transliteration with Latin characters (Romaji input method); the other uses Japanese characters (Kana input method). Although 70% of native Japanese speakers utilize the Romaji input method, users expect to have a choice between the two.

For input of Indian languages, the process is simpler but still involves a variety of input methods. Even for the PC, there are number of non-standard keyboard layouts and encoding. At least three types of keyboard systems are in use today for Indic text input:

-- phonetic input of Indic characters
-- transliteration of Indic characters using a keyboard with Roman script
-- old typewriter layouts

With the exception of Tamil, no standard has been developed by Governmental or other bodies to govern input of other Indic languages. The inscrypt layouts of Department of Electronics, India, used in iLeap and some other software, are considered the de facto standards for pan-Indic software developers but for other Net Appliances, no informal or formal standard exists.

LARGE CHARACTER SET, SMALL KEYPAD

Beyond the basic functions necessitated by Asian languages, there are other ways to improve the performance of the User Interface and enhance the end-users text entry experience. When these additional functions are combined with the basic functions the UI can be described as intelligent. As Internet use and Internet appliances proliferate so will the demand for intelligent UI. An intelligent UI makes entering text on an Internet appliance faster and easier, in any language.

Ideographic languages are often made up of thousands of characters, which make inputting text quite complex - there are many more characters than there are keys on a keyboard. This means that each key on the keyboard cannot possibly represent 1 or 2 or even 3 characters; multiple keystrokes must be used to represent a character or series of characters. In Japanese, for example 46 phonetic symbols, 2 diacritical marks and 7 punctuation marks are combined to create thousands of possible character combinations. A commonly used keyboard mapping found on mobile phones in Japan uses 10 digits to represent the 46 phonetic symbols and 2 digits to denote the 2 diacritic marks and 7 punctuation marks. To input the frequently used word "Mono", you would find yourself using 10 keystrokes to input this 2-syllable word. Cutting-edge technology, such as Slangsoft intelligent Text Input and Display Platform (iTID), enables this same word to be entered with only 2 keystrokes, by utilizing linguistic and frequency of usage data to suggest a word. It can also offer word completion, as well as a list of word options (mane, mune, mine, etcetera).

Most of the technology that is available today to enable fast input of text from Internet employs linguistic dictionaries. However, since maintaining a super small footprint is critical to any embedded device and most Net appliances, efforts are constantly made to increase efficiency of fast text input by offering more functionality with smaller RAM requirements. Slangsoft's iTID Platform, for example, includes linguistic data with frequency of usage which relies on a proprietary algorithm which combines root and word inflection rules to greatly expand extremely compact wordlists for each language to keep the size of the linguistic data very small.

A LOOK TOWARD THE FUTURE

The future holds the promise that computing capabilities will be contained in everything that human beings interact with such as cars, furniture, and clothing. The manufacturers of these items, along with the makers of applications that run on them, will be challenged to provide fast, intuitive ways for people to interact with them.

The introduction of personal computing was the biggest event in the computing world in the last 20 years and has proven to be an unstoppable phenomenon that is evidenced today in ubiquitous computing. People access information on the go with mobile phones, PDAs and pagers and at home with set top boxes and MP3 players. In the future, personal computing will be available everywhere, in cars, elevators, eyeglasses, and more, changing the shape of our actions and interactions as we know them today. Keys will become obsolete, for example, as doors are embedded with software that identifies your handprint to open while displaying your recent email messages.

Slangsoft remains committed to its mission to enable fast, intuitive interaction between people and the devices they use, now and into the future, to promote a cost-effective economy. We applaud the work of INFITT and the Tamil Internet conference in furthering this goal by facilitating the widespread use of Tamil in conjunction with information appliances and applications.

We used to say: Engum Tamil, Eithilum Tamil. (Tamil Everywhere, Tamil in anything). In the era of internet, how do you interpret this statement? By introduing Tamil in all communication/information devices and systems. By making it fit for ubiquitous computing. This is essence or our vision.

We thank you for the opportunity to appear before you to express our views and look forward to a continued, productive partnership.