LICT4D.pk

Dr. Sarmad Hussain and Sana Gul
Center for Research in Urdu Language Processing (CRULP)
National University of Computers and Emerging Sciences (NUCES)
sarmad.hussain@nu.edu.pk

INTRODUCTION

Information has now become such an integral part of our global society, that its access is considered as a basic human right [1, 2]. Moreover, progress of rural and urban developing populations is also getting increasingly dependent upon access to information [3]. This is specifically applicable to Pakistan, which is a developing country having GDP per capita about US$430, 67.7% rural population and a third of population below the poverty line [4, 8]. Access to information is thus critical for Pakistanis.

Information and Communication Technology (ICT) has emerged as the dominant carrier of information across the globe and therefore access to information can be equated with access to ICT, specifically the Internet. Currently, 1.7 million Pakistanis use the internet, which is only 1.2% of the total population of Pakistan [8]. This shows that there is enormous potential for Internet usage in Pakistan, and therefore an immense opportunity for growth through access to information through the Internet. Investments have been put into developing ICT infrastructures in Pakistan. Under the previous and current regimes, the number of cities which have access to the internet has increased from 400 in 2001 to more than 1050 in 2003 [8, 9,10]. Nevertheless, the persisting digital divide attests that the current path towards providing connectivity and technology infrastructure alone would not enable the majority of Pakistani population to benefit from the present information availability [5].

Pakistan is culturally and linguistically a diverse country. There are 57 languages spoken in Pakistan [12]. Only about 5-10% [8,11] of these people can communicate in English, majority of whom reside in urban areas. Urdu, Punjabi, Sindhi, Baluchi, Pashto and Saraiki are the most commonly spoken languages in Pakistan. Pakistan has about 50.5% adult national language literacy rate according to one estimate and the language-wise population distribution of Pakistan is as follows [8]:

Language	**Punjabi**	**Sindhi**	**Siraiki**	**Pashtu**	**Baluchi**
Total Speakers	45,000,000	17,002,000	30,000,000	9,585,000	5,681,000
Percentage of total population of Pakistan	30.3%	11.5%	9.8%	8.4%	3.83%

These figures clearly indicate the vast linguistic diversity of the country. Currently, English is the lingua franca for ICT, with majority of content being in English [6]. This makes English language information available on ICT inaccessible to a more than 90% majority of Pakistanis. This is reaffirmed by Digital Review of Asia Pacific: "The ready availability of relevant local language content is critical for the development of productive capacity … one of the challenges … of Internet diffusion [lies] in the dominance of the English language … content … with little relevance to many Asia-Pacific Internet users" [7].

Further as English is the official language of Pakistan, there is a large amount of English language content generated by government of Pakistan, which has much relevance to its citizens who do not speak English. Therefore, it is equally important for e-governance to localize this information.

Localization of ICT is as critical as the infrastructure for development. This is being termed as LICT4D (Localized Information and Communication Technologies for Development).

CHALLENGES FOR LOCALIZATION

Pakistani languages are very different and in some aspects much more complex than English, e.g. Pakistani languages have a very complex writing system. The current technology does not completely support the requirements of these local languages, as it was initially developed to support only English. Thus, current technology needs to be developed and improved to support Pakistani languages. Second, there is shortage of human resource in Pakistan skilled to enhance the existing technology to support these local languages. Finally, there are very few studies which look at local language problems to devise and recommend effective policy framework to address problems and devise short-term and long-term solutions.

Enabling local language computing poses a unique challenge to English and western-centric ICT. However, attempting to develop local language ICT would initially require the development of certain local language standards. These standards include character sets for Pakistani languages on national and international fora (e.g. Unicode, ISO/JTC SC2/WG2), keyboard and keypad layouts, normalization and sorting sequences, and terminology translation to make computing interface available in local languages. The scope of standardization may also contain interfaces for mobile and other hand-held devices.

In addition to definition of standards, work also needs to be done in development of technology. Technology needs to be developed in the areas of script, speech and language processing. Within these areas, there are large variety of application development required, some being more core and fundamental in nature, while other more advanced, complex but equally vital for end users. These applications include Keyboard driver, Normalization and Sorting Utility, Find/Replace Utility, Fonts, Handheld device fonts, Lexicon, Thesaurus, Natural Language Processor, Spell Checker, Grammar Checker, Text Corpus Development for NLP Applications, Text to Speech System, Speech Recognition System, Machine Translation System and Optical Character Recognition System. Brief explanations of these applications are given in Appendix A. These applications may be built on Microsoft or Linux platforms. For cost-effective access, Linux software should be developed. However, before providing specific software, basic, light-weight and robust Linux distributions must be created in local languages. Status of standards and local language applications for local languages of Pakistan is given in Table 1 below.

Table 1: Status of standards and applications for local language of Pakistan

x → *completed* *xx* → *work done outside Pakistan*
*** → *significant progress* *-* → *some work done*

	Urdu	Punjabi	Sindhi	Pashto	Baluchi	Saraiki
Standards						
Character Set	**x**	**-**	**x**	*****	*****	
Keyboard	**-**		**x**	**xx**		
Keypad (Telephone)						
Collation Sequence	**-**		*****			
Interface Terminology	*****			**-**		
Handheld Device Interface Terminology						
Locale	**-**			**xx**		
Software						

Applications						
Keyboard driver						
Normalization and Sorting Utility						
Find/Replace Utility						
Fonts	**x**	**-**	*****	*****	**-**	
Mobile/PDA Fonts						
Lexicon	**-**					
Thesaurus						
Natural Language Processor	**-**					
Spell Checker	**-**					
Grammar Checker	**-**					
Text Corpus Development for NLP Applications	**-**					
Text to Speech System	**-**					
Speech Recognition System	**-**					
Machine Translation System	**-**					
Optical Character Recognition System	**-**					
Linux Distribution	**-**					

The table shows that the matrix is very sparsely filled. There is hardly any work done in the Pakistani local languages which makes Pakistan a fertile ground for R&D in linguistics, speech processing, computational linguistics and computer science.

RECOMMENDATIONS

Academia

The educational institutions of the country should play the leading role in R&D in local language computing. They should look into the inter-disciplinary challenges and conduct original research in applying exiting solutions for Pakistani languages or develop better and more optimum computational models tailored to the specific needs of these languages. They should also look into other fundamental or auxiliary programs. It is appalling to note that with so many languages spoken in the country there was no linguistics department in any university in Pakistan as late as 1990's. Even today, there is hardly a couple of very basic programs. Additionally, there are only one formal computational linguistics degree programs in Pakistan.

This research should be institutionalized and made sustainable by developing research centers or research groups in localization and involving interested students who can prototype innovative ideas in related areas.

Academia form marketing and social science should also research into effective marketing and dissemination strategies for the technology being developed to put it into effective use by end-users for their development. They should especially look into effective use of technology, supported by localized interface, by rural populations.

Academia can also play a significant role in the development of HR capacity for private commercial and social sector in these areas.

Government

Governments can play the most crucial role in the promotion of local language computing. Increasingly governments, even of developed countries, are becoming aware of the need for localized information. Effective policy-making is critical and essential for the development of localization. This can be achieved through multi-pronged approach. For example, by imposing restrictions on sale of software in Spain which did not support both Spanish and Basque, Spanish government automatically convinced vendors like Microsoft to support these languages in its products.

Equally effect way of promoting local language computing is by creating the demand. For example, requiring multi-lingual websites by federal, provincial and district governments may generate a significant interest in R&D in this area. This would also benefit the masses, who will get access to critical information.

Government can also play a vital role by funding HR development in this area. HR support may focus on higher education, research and development promoting local language computing and related areas.

Another major issue which has been a major obstacle in the development of local language computing has been lack of protection of Intellectual Property of the people doing R&D. Protection of copyrights is an important issue, and has been the primary cause of demise of the word-processing development industry in Pakistan which flourished in 1980's. Enthusiasts invested time and resources in development of excellent software for DTP in local languages. However, as most could not reap the benefits of their hard work, all these initiatives were eventually shelved in 1990's.

Social Sector

The social sector organizations, including the non-governmental organizations (NGOs), can be instrumental dissemination of local language technology and end-user training to put it to effective use.

These organizations can also play a significant role in defining the need for having the local language solutions. By studying the rural populations and poor urban populations of the country and gathering data on the real life issues they can identify the need of various types of applications and content required.

Private Sector

The private sector is mainly responsible for leveraging the research and prototyping by the academia and translating it into commercial grade localization applications. This would also effectively utilize the trained human resources and foster meaningful links with academia.

CONCLUSION

As quoted above most population of Pakistan require urgent access to information for their development. Internet is the main repository and vehicle for access of information. However, language barriers pose a significant problem to this access. This is an urgent and important issue and must be addressed to give information access to poor and illiterate populations of Pakistan, the access which is their basic human right! This may be effectively achieved by computer and social scientists in public sector, academia and private sector working closely together to develop language technology and effective dissemination mechanisms for this technology. This is a challenging inter-disciplinary task and a frontier of Information Technology.

REFERENCES

[1] "**Info-structure in developing countries,**" UNESCO's contribution for the World Summit on the Information Society (WSIS) (8 February 2002).

[2] "**Statement Presented by Minister of Information and Communications Technology of Thailand,"** Regional Conference for WSIS in Tokyo Jan 12-15, 2002).

[3] **"Access Features Information Technology for Rural Community Development**," http://www.ncsa.uiuc.edu/News/Access/Stories/InfoTechnology/information.html

[4] http://www.worldbank.org/data/countryclass/classgroups.htm#top.

[5] "**Information and Communication Technology in Developing Countries of Asia**," Brahm Prakash, http://www.adb.org/Documents/Conference/Technology_Poverty_AP/adb6.pdf

[6] "**The effect of native language on Internet usage**", Telecommunications Policy Research Conference (TPRC) 29th Research Conference on Communication, Information and Internet Policy, October 27-29, 2001, Alexandria, Virginia.

[7] "**Content and ICTs: challenges, innovation and prospects**" Digital Review of Asia Pacific 2003/2004, www.digital-review.org.

[8] "**Pakistan**" Digital Review of Asia Pacific 2003/2004, www.digital-review.org

[9] **"Development of IT"**, http://www.dawn.com/2001/01/29/letted.htm#1

[10] "**560 towns have Internet access, says Dr Atta**", http://www.dawn.com/2002/02/03/nat6.htm

[11] **"Languages of Pakistan",** http://www.ethnologue.com/show_country.asp?name=Pakistan

[12] "**Language ideology and Power: Language learning among the Muslims of Pakistan and North India**", Tariq Rahman, Oxford University Press, Karachi, Pakistan

APPENDIX A: Briefs on Standards and Applications for Localization of Computing

Localization Standards

Character Encoding

There are a variety of characters and scripts with which different languages of the world are written. To enable writing these characters within a computer or other ICT devices (which only code numbers) each of these characters has to be given a unique number. This encoding must be standardized so that local language computing can produce data which is exchangeable among various applications. The only international standard, also adopted by ISO, is ISO IEC 10646 commonly known as Unicode character encoding standard. This standard is designed to support all languages of the world, and is under constant development. Many languages of the world are still not completely supported.

Keyboard Layout

Once the characters have been encoded, keyboards or Keypads (on smaller devices) offer a standard way of inputting text in different languages. However, different languages with different sets of characters require different keyboard layouts to be defined. Computers pose more

requirements than typewriters, so existing older versions of keyboard layouts need to be updated. The keyboard layout may be designed on basis of frequency analysis, ad hoc standards, or a combination of both. Vendors like Microsoft would only support a keyboard within their software if they have been standardized by respective populations and government(s).

Collation Sequence

Any significant data processing requires data sorting. Sorting is normally language and culture sensitive and must be defined clearly. This sorting definition is called collation sequence of characters for a language. Many languages of the world still need to define a crisp collation sequence, as the ones used for dictionaries are sometimes vague as they have been developed for human users, who do not require exact definitions like computers.

Interface Terminology

Even when local language computing is possible, sometimes it becomes difficulty to operate the ICT devices (like computers and PDAs etc.) because their interface is in English. Thus, the standard terminology used by these interfaces (e.g. vocabulary in the computer menus: File, Edit, Save, Save As, Exit, Insert etc.) must be translated into local languages in a coherent manner so that all devices start using same local language terminology. There are no international organization(s) which manage this and must be done at national level.

Locale

Locales define the way information is displayed using ICTs for a particular country. For example date formats are different for UK and US, Europeans use comma to indicate fraction part of a number but Americans and British use a decimal point. These need to be standardized by each country.

Localized Applications

Keyboard Driver

Keyboard Driver is the computer program which maps the different keyboard keys onto the characters for any language. Thus, this program implements the keyboard layout defined for a language.

Normalization and Sorting Utility

Sorting Utility implements a sorting algorithm (collation sequence) for a language. However, sometimes the text needs to be processed before it can be sorted, depending on the language and the encoding scheme being utilized. For example, for character Hamza in Urdu there are two Unicode points, which must be equated before they are sorted, as both must sort the same way. This is achieved by Normalization software applications.

Find/Replace Utility

This is a commonly used utility in desktop publishing, used to find a text string and replace it with another one. However, searching and replacing text strings may involve differing algorithms across languages and thus this utility has to be defined separately for these languages. For example, English has small and capital letters which may or may not be searched separately, but languages using Arabic script do not have this concept of capitalization.

Fonts

Fonts are character representations for a language in a form which a computer or other ICT devices can display. There are multiple font definition formats for different vendors like Microsoft, Apple, etc. Also the font standards vary depending on the ICTs, e.g., fonts for computers may not work for mobile devices, etc.

Lexicon

Lexicon represents a dictionary for a language for human and software application consumption. Humans users use language lexica for conventional use, e.g. in desktop publishing. However, these lexica are also used by other computer applications like Spell Checker, Machine Translation, etc. Lexica for these applications may become very advanced, extensive and abstract.

Thesaurus

Thesaurus suggests synonyms and antonyms of words and is now normally used in desktop publishing to generate content on ICTs.

Natural Language Processor

Different languages pose different challenges to technology. For example, some languages do not put a space between different words. Thus, it would not be possible for the computer program to determine where to break a sequence of characters at the end of a line. Humans can do it because they use a mental lexicon. Such (language dependent) applications come within the scope of Natural Language Processor.

Spell Checker

Spell Checker is used to check spelling in desktop publishing applications. It normally requires a Lexicon.

Grammar Checker

Grammar Checker is used to check grammar in desktop publishing applications. It normally requires a Lexicon.

Text Corpus Development for NLP Applications

Advanced analysis in Natural Language Processing (NLP) normally requires a large amount of text in the respective language. The analysis could be on frequency of usage of different words in various contexts (e.g. N-Gram Analysis), simple frequency analysis, etc. However, to get reasonable statistical information about a language a balanced and large amount of text is required. Corpus may contain tens of millions of words.

Text to Speech System

Text to Speech system takes regular text in a language from a file or keyboard as input and converts this text to speech in that language using linguistic knowledge and digital signal processing techniques. TTS system is especially useful tool to access content for blind and illiterate people.

Speech Recognition System

Speech Recognition system is basically a dictation system, which can enable computer to type what the user is speaking. Speech recognition systems are mostly speaker dependent (for higher accuracy) and therefore must be trained before they can be used. They are very good tools for content creation, especially for illiterate populations.

Machine Translation System

Machine Translation system translates text from source language to target language automatically using grammars of both languages involved. Currently, though one could get a grammatically correct translation, meaningful (semantically correct) translations are hard to get automatically and human quality assurance is normally required. However, research and development in this area is improving the quality of MT. Even rudimentary MTs can bring the enormous English and Spanish content already available on the net to local populations, the former being the source languages and the latter being the target languages.

Optical Character Recognition System

OCR converts scanned documents into text automatically. Thus, instead of typing up records and books, one could just scan and process it through OCR to create online content very quickly and effectively. OCR normally requires advanced image processing and is potentially a great tool for content creation.

Linux Distribution

Linux Distribution means a release of Linux which will compile and run. The distribution may have many utilities and packages, but a minimal distribution from a normal computer user's point of view would still minimally contain desktop publishing tool(s), email and internet client and printing facility. To make the distribution effective for use by non-English speakers, the minimal package should support local language script and the interface must also be in local language.