

Multilingual ICT Education: Language Observatory as a Monitoring Instrument

Mohd Zaidi Abd Rozan¹, Yoshiki Mikami¹, Ahmad Zaki Abu Bakar², Om Vikas³

¹Nagaoka University of Technology, Nagaoka, Niigata, JAPAN 940-2188

²Universiti Teknologi Malaysia, 81310 UTM, Johor, MALAYSIA

³Department of Information Technology, New Delhi, 110 003 INDIA

zaidi@mis.nagaokaut.ac.jp

mikami@kjs.nagaokaut.ac.jp

zaki@fsksm.utm.my

omvikas@mit.gov.in

Abstract

Ubiquitous learning challenges students to become adept at information retrieval, management and synthesis from a variety of sources. This sparks discovery activities that are student-centred and personalized. Personalized means that the learning is best conducted in the natural language of the student. Language is an important tool for human communication and at the moment, the language dominating ICT is English. Although many efforts have been made, learning English is a slow and expensive process. There were also many casualties and sacrifices which unfortunately were at the expenses of many local and indigenous languages and cultural heritage. This paper presents an effort made by a consortium of universities and research centres in Asia to address the problem of language digital divide by establishing a World Language Observatory. Compared to an astronomical observatory, which observes space for astronomical phenomena, a language observatory observes language phenomena in cyberspace. Software agents in the form of soft bots are periodically sent into cyberspace by the mother Language Observatory in Japan to examine websites and identify its languages and contents in an attempt to identify language communities in various regions of cyberspace. Assisted by various language observatories around the world a language census chart is then published annually on the UNESCO's International Mother Language Day to inform the world of the current language situation in cyberspace which have implications on education.

Keywords: ICT education, Ubiquitous learning, Digital divide, Multilingualism, Language, character script, standardization, Language Observatory Project.

Copyright © 2005, Australian Computer Society, Inc. This paper appeared at *South East Asia Regional Computer Confederation (SEARCC) 2005: ICT Building Bridges Conference*, Sydney, Australia, September, 2005. Conferences in Research and Practice in Information Technology, Vol. 46. Graham Low, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

1 Emerging ICT Education Issue

Just as the industrial era of mass production had now given way to more customized manufacturing in the knowledge-based economy, our educational institutions are facing the daunting challenge of moving away from mass education toward customized learning to be offered at any place, at any time and with any means. The current situation demands educational institutions find ways to develop learning skills required by a 21st century workforce; particularly ready-to-be-employed skilled workers with an ability to innovate and who could learn how to learn. This emerging issue of how our current educational systems could change to foster an environment of customized learning among students has prompted many to seek Information Communication Technology (ICT) solutions.

Mobile or ubiquitous learning is one of the learning models becoming more popular recently and superseding e-learning. This new learning technology came into being through the advent of affordable wireless computer-based information devices brought in by the pervasive use of the Internet and the convergence of multimedia technology and the wireless application protocol (WAP). According to Goldberg (2002), in a ubiquitous model, students in schools and campuses must become adept at information retrieval, management, and synthesis, from a variety of sources. Personal technology puts those resources within their reach, not just in the hands of the teacher, librarian, or lab aide. The ubiquitous model gives students the means of communicating and requires them to develop responsible "netizenship", with guidance from the instructor. This model is also student-centred and personalized, based on discovery activities. It is both collaborative and self-directed. Personalized education also means the learning is best administered in the natural language of the student.

Although this model is very pervasive and the technology is superb, we are still confronted with an aged old problem which we have not been able to successfully eradicate. This problem relates to the issue of "digital divide" or "e-exclusion". While the benefits of ICT are many, the negative outcomes as circumscribed in what is universally termed as the "digital divide" grows wider and is causing grave concern. A nation's competitiveness is tied to its capacity for ICT creation and application. If the disparity

in wealth divides the rich and the poor, and the disparity in education divides the literate and non literate, then the digital divide, refers to the disparity between those who have use of and access to ICT versus those who do not. Digital divides exist both within countries and regions and between countries. It transcends locality, races, gender, age, languages, culture and religion. The issue of the digital divide is more than direct access to technology, it is also regarding the disparity between how different nations are using ICT as a tool for social and economic development. However, this paper will focus more on the language-related issue. Language is an important tool for human communication and at the moment, the language dominating ICT is the English language.

2 Diversity of Languages and Scripts

Customized education has to cope with the tremendous diversity of world languages and scripts. Language experts estimate that today nearly 7,000 languages are living languages on the globe¹. In terms of official languages, the number of languages is still large and could be more than three hundred. The United Nations Higher Commission for Human Rights (UNHCHR) has translated a text of

universal value, the Universal Declaration of Human Rights (UDHR), into as many as 328 different languages.

These translated texts can be viewed by visiting the UNHCHR website². Among all the languages appearing in this site, Chinese has the biggest speaking population of almost a billion people, and is followed by English, Russian, Arabic, Spanish, Bengali, Hindi, Portuguese, Indonesian and Japanese. The language list continues until those languages with less than a hundred thousand speakers.

The site also provides the estimated speaking population of each language. When we sort out languages by speaking population as the key and plot each language in a logarithmic scale chart, then the relationship between speaker population and its ranking emerges as something like a Zipf's-Law curve as shown in Figure 1 with at least at range between ten to hundredth. This means that the tenth language (in this case, Japanese) has one tenth of the speaking population of the top ranked language (in this case, Chinese) and the hundredth language (in this case, Turkmen) has one hundredth of the top language.

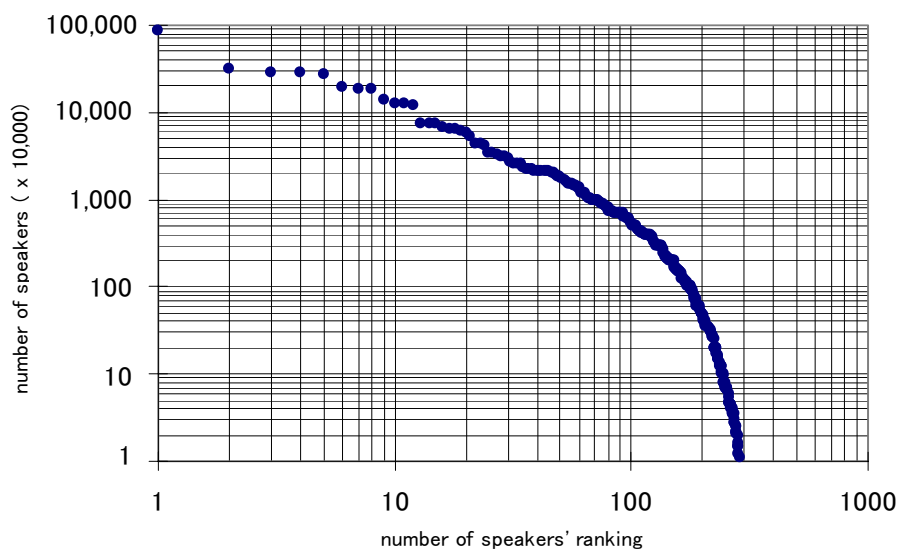


Figure 1: Quasi Zipf's Law Curve of Language Speakers

From the viewpoint of complexity in localization, diversity of scripts is another and a more problematic issue. "How many scripts are used in the world" is a difficult question to answer. It depends on granule size of counting. In this paper, for the sake of simplicity, we treat all Latin based scripts, alphabets plus its extensions used for various European languages, Vietnamese, Filipino, etc. as one set. We will also treat languages using Cyrillic scripts as one set and so on for languages based on the Arabic script. In the same nature, we will treat Chinese ideograms, Japanese syllabics and Korean Hangul scripts as one. The remaining languages will comprised of many kinds of

diversified scripts. Here, we will take the "Indic script" to be in the fifth category. This category includes not only Indian language scripts such as Devanagari, Bengali, Tamil, Gujarati, etc. but also four Southeast Asian language scripts; Thai, Lao, Cambodian (Khmer) and Myanmar. In spite of the differences in their shapes, these scripts have the same origin (the ancient Brahmi script) and have the same type of behaviours in formulation. When we summed up the speaking population of each language by this script grouping, the number of users of each script is summarized in Table 1.

Script	Latin	Cyrillic	Arabic	Hanzi	Indic	Others
Number of users in million	2,238	451	462	1,085	807	129
[% of total]	[43.28%]	[8.71%]	[8.93%]	[20.98%]	[15.61%]	[2.49%]

Source: Speaking population of each language is based on the data provided at UNHCHR website².

Table 1: Distribution of User Population by Script Groupings

3 Education and the Mother Tongue

As pointed out earlier, English is the language with the second largest speaking population. According to the UDHR website, the number of persons speaking English as their mother tongue is 322 millions. This represents 6.2% of the global population, although it is possible this statistic represents the situation a decade ago. Nevertheless, a market research company estimates that as many as 37% of Internet users are English speakers³. Another study (O'Neill, 2003) found a higher proportion of English usage to be 72% in terms of web pages which were recorded by analysing random samples of web pages.

The high rate of English usage in cyberspace can be explained through various ways. One main reason is due to the fact that English is a dominant world language. Globalization of economy requires a *lingua franca* for market participants and for this purpose, English is chosen for business communication. Many Information Communication Technology products also originate from English speaking countries and as such it is quite natural for the systems and their documentation to be in English. Another contributor could be due to the fact that computer networks which span the world necessitate a medium of communication. As such, it is quite natural for English to be one of the important components of the protocol set in the presentation layer of the communication model.

Although English is acting as the *lingua franca* in the business world and on the Internet, it has also many limitations pertaining to education. There are certainly many merits for using a single *de facto* language like English but studies have shown that, in many cases, instruction in a mother tongue is more beneficial for students in regards to acquisition of language competencies, achievements in other subject areas, and even for learning a second language⁴.

A recent study on E-Learning initiatives in India by Ravichandran (2005) has found out that a special but biggest problem faced for the implementation of e-Education lies on its language medium. As the population mainly converse in their mother tongue, this has produced a powerful effect to the hampering of on-line education which are mostly in English. Looking from another perspective, the dominance of English in most on-line education is another worrying scenario. Since the coverage is also targeting to primary school level, this would be an obstacle to the usage of mother language in its cyberspace.

Only less than 10% of computers in Sri Lanka use Sinhala and Tamil (APDIP, 2003). The main operations are mostly for word processing and publishing and sadly insignificant usage in local languages. With such a low usage in mother

language, it is likely that the competitive nature of English language will dominate and supersede the mother language in Sri Lankan cyberspace.

A survey of the Internet user profile compiled by the Thai government⁵ provides us an interesting statistic in this context. Its cyberspace population as in 2004 is reaching 7 Million users. With around 12% of the users grouped in the last two ranks (limited and no proficiency in English), it seems that this group of users will be at the losing end as shown in *Table 2*. It means that concentration to English in the global pool of knowledge will lead and enhance the divide in access to information in a global scale.

Excellent	Good	Fair	Limited	None	Total
7	38	42.9	11.3	0.8	100%

Source: National Electronics and Computer Technology Center (NECTEC), Thailand, Internet User Profile of Thailand 2004⁵.

Table 2: English Proficiency and the Internet Usage Rate

Not restricted to such information, our latest observatorial analysis found that there are 4332 web servers with subdomains of .ac and .edu in Asian country code Top Level Domains (ccTLDs). This contributed to more than one fifth of the total with an estimation of nearly 10 millions in text documents. By means of such info structure, it is mainly important to ensure that there is room for the usage of mother languages for their very existence.

4 ICT and Multilingualism

Compared to a decade ago, current ICT systems are capable of handling multilingualism to a certain degree. Thanks to the emergence of a standard for multilingual character coding in the form of the ISO/IEC 10646 standard which is also used for the Unicode standard, as well as the sophisticated internationalization of software carried out at various levels, the number of languages being supported by ICT for the last decade have increased. Although many efforts have been made for the localization of major platform/application software by vendors, the language coverage of these softwares is still limited. The most recent version of Windows XP (Professional SP2) is able to handle a long list of 123 languages now available to users. However, if we look at the list more closely, most of the languages are for European languages and very few of are for Asian and African languages. The language coverage is summarized in *Table 3*. In this table, languages are categorized by the grouping introduced in the first section of this paper. Hence, the population-based coverage of Windows XP is calculated to be around

83.72% against the global population. This is not a bad figure, but as we will discussed later in this paper, this figure seems to be an overestimated figure which does not tally with reality.

Script Region	Latin	Cyril	Arabic	Hanzi	Indic	Other
Europe	European*	Russian, Macedonian & Slavic languages**	---	---	----	Greek Georgian Armenian
Asia-Africa	Afrikaans Azeri English Vietnamese Malay Indonesian Swahili Tswana Uzbek Xhosa Zulu Turkish	Mongolian Azeri Kazakh Kyrgyz Uzbek	Arabic Urdu Persian	Chinese Japanese Korean	Gujarat Tamil Telugu Kannada Bengali Malayalam Punjabi Hindi Marathi Sanskrit Konkani Thai	Assyrian Dhivehi Hebrew

*Includes: Albanian, Basque, Bosnia, Catalan, Croatian, Czech, Danish, Dutch, English, Estonian, Faroese, Finnish, French, Galician, German, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Sami, Serbian, Slovak, Slovenian, Spanish, Swedish and Welsh.

**Includes: Belarusian, Bulgarian, Serbian, Bosnian & Ukrainian.

Table 3: Windows XP SP 2 Coverage on Language by Major Script Categories

Beside language scripts, search engines are indispensable components of the global information society. Vast pool of knowledge can be made accessible through the function of search engines. When we investigate the language coverage of many popular search engines, the situation is far worse compared to the case of the Windows availability. One of the globally used multilingual search engine, Google, is found as of April 2005, to have indexed more than 8 billion pages written in various languages. However, the languages covered so far is limited to only

some 35 languages. Among these, the Asian languages covered by Google are only eight, i.e. Indonesian, Arabic, Chinese Traditional, Chinese Simplified, Japanese, Korean, Hebrew and Turkish (see *Table 4*). Among the major languages of the SEARCC region not included in Google's coverage are all the Indian languages, Urdu and Sinhala. If we calculate the population-based coverage, it will decrease to 61.37% largely because Asian and African language pages are not searchable.

Script Region	Latin	Cyril	Arabic	Hanzi	Indic	Other
Europe	European*	Russian Bulgarian Serbian	---	---	---	Greek
Asia-Africa	Indonesian		Arabic	Chinese (Traditional & Simplified) Japanese Korean		Hebrew Turkish

*Includes: Catalan, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish.

Table 4: Google Coverage on Language by Major Script Categories

As mentioned in the first section of the paper, if we visit the website of the Office of the Higher Commissioner for Human Rights of the United Nations, we will find more than 300 different language versions of the Universal Declaration of Human Rights (UDHR) starting from Abkhaz and ending with Zulu. Unfortunately, we will also

find many of the language translations, especially for non-Latin script based languages, are just posted as “GIF” or “PDF” files and not in encoded texts. We again summarized the situation by the script grouping as matching as the previous tables (see *Table 5*).

Form of Presentation \ Script	Script					
	Latin	Cyril	Arabic	Hanzi	Indic	Others
Encoded	253	10	1	3	0	1
PDF	2	4	2	0	7	10
Image (GIF)	1	3	7	0	12	7
Not available	0	0	0	0	3*	1**

*Magahi**, *Bhojpuri**, *Sanskrit** and *Tigrigna***.

Table 5: Form of Representation of the UDHR Texts by Script Grouping

The table clearly shows that languages which use Latin scripts are mostly represented in the form of encoded texts. Languages which use non-Latin script especially Indic and other scripts on the other hand, are difficult to be represented in encoded form. When the script is not represented by any of the three foremost forms provided, they are grouped as not available. Moreover, it is compulsory to download special fonts to properly view these scripts. This difficult situation can be described as a digital divide among languages or termed as the “language digital divide”.

5 Character Coding Standards for a Multilingual Cyberspace

Technology brought to a zone that is culturally, environmentally and socially different from its originating source will in no doubt face many challenges. From a technical viewpoint, the major reason behind the language digital divide is due to the lack or non-availability of appropriate character encoding schemes. Due to this fact, the UDHR website creators have to put text not able to be encoded but in the form of PDF or images. If we look at internationally recognized directories of encoding schemes, like the IANA Registry of character codes6 or ISO-IR7 (International Registry of Escape Sequence), we can not find any encoding schemes for these languages which we termed as have fallen through the net.

A question may arise from this situation. Now that major computer platforms like Microsoft Windows are providing solutions for these languages, why is it then that we must still suffer from such a digital divide? Interestingly, the languages supported by Windows are not substantial especially to cover minority languages. Again, Table 3 provides relevant details on this issue.

Character coding standards that are recognized by International organizations such as IANA or ISO-IR are standards that were created through a top-down approach. Unicode for example, provides character encoding schemes for 50 writing systems from English to Osmanya and through Kannada. Unicode with its latest version 4.1.0

covers a vast system of encoding properties. Unfortunately some scripts are not exhaustively supported, such as Balinese, Javanese, etc. According to Narayanan (2004), local vendors in India were complaining about the lack of tools and technical expertise to implement Unicode. Another contributing factor for the lack of supported scripts could be due to the small user communities coming from very small economies who do not have a strong lobbying capability. These languages are not likely to have their scripts included in the standard and hence not implemented on the web. In a short while, languages like these will most likely disappear and their cultural heritage lost. This situation has motivated some initiatives to be launched by mostly local vendors and academics to save the near extinct languages. However, lack of resources and know how, limits their successes.

It is also important to note that many character encoding standards that were established at the national level are also present for many languages. These standards are identified as National Standard (see Table 6). In the case of the family of Indian writing systems, the first national Indian standard was announced in 1983. It was named the Indian Standard Script Code for the Information Interchange (ISSCII). Later in 1991, it was amended to become the second version, national standard IS 13194 and is referred with a slightly different abbreviation as ISCII which is currently in use in India. However, although there exists a national standard, hardware vendors, font developers and even end-users have been creating their own character code tables which lead to a chaotic situation. The creations of so called exotic encoding scheme or local internal encoding have been accelerated particularly through the introduction of user-friendly font development tools. Although the application systems working in these areas are not stand-alone systems and are published widely via the web, the necessity for standardization has not been given serious attentions by users, vendors and font developers. The non-existence of professional association and government standard bodies is another reason for this chaotic situation.

Economies	Language	Writing System	National Standard (First version)	National Standard (Current version)
Bangladesh	Bengali	Bengali	BDS 1520:1995	BDS 1520:2000
Brunei	Malay English	Latin Latin	- -	- -
Bhutan	Dzongkha	Dzongkha	-	-
Cambodia	Khmer	Khmer	-	-
China	Chinese Mongolian Uighur Kazakh Korean Yi	Hanzi (simplified) Mongolian Uighur Kazakh Korean Yi	GB 2312:1980 GB 8045: 1987 GB 12050: 1989 - GB 12052:1989 GB 13134: 1991	GB 2312:1980 GB 8045: 1987 GB 12050: 1989 - GB 12052:1989 GB 13134: 1991
India	English Hindi/Konkani/ Marathi/ Nepali/Sanskrit Punjabi Gujarati Oriya Bengali Assamese Telugu Kannada/Konkani Malayalam Tamil Urdhu/Sindhi/Kashmiri	Latin Devanagari Punjabi/Gurmukhi Gujarati Oriya Bengali Assamese Telugu Kannada Malayalam Tamil Perso-Arabic or Devanagari	ISSCII: 1983 ISSCII: 1983 ISSCII: 1983 ISSCII: 1983 ISSCII: 1983 ISSCII: 1983 ISSCII: 1983 ISSCII: 1983 ISSCII: 1983 ISSCII: 1983 ISSCII: 1983	IS13194:1991 (ISCII) IS 13194:1991 (ISCII) IS 13194:1991 (ISCII) IS 13194:1991 (ISCII) IS 13194:1991 (ISCII) IS 13194:1991 (ISCII) IS 13194:1991 (ISCII) IS 13194:1991 (ISCII) IS 13194:1991 (ISCII) IS 13194:1991 (ISCII) IS 13194:1991 (ISCII)
Indonesia	Indonesian	Latin	-	-
Japan	Japanese	Kana Kana + Kanji	JIS C6220: 1969 JIS C6226: 1978	JIS X0201: 1997 JIS X0208: 1997 JIS X0212: 1990
Rep. of Korea	Korean	Hangul + Hanja	KS C5601: 1987	KS X1001: 1992
DPR Korea	Korean	Hangul + Hanja	KPS 9566: 1997	KPS 9566: 1997
Laos	Lao	Lao	-	-
Malaysia	Malay	Latin Jawi	- MS 1368: 1983	- MS 1368: 1994
Mongolia	Mongolian	Cyril + ext. Mongolian	-	-
Myanmar	Myanmar	Myanmar	-	-
Nepal	Nepali	Nepali	-	-
Pakistan	Urdu	Perso-Arabic	-	-
Philippines	Filipino English	Latin + ext. Latin	- -	- -
Singapore	English Chinese Malay Tamil	Latin Chinese Latin Tamil	- - - -	- - - -
Sri Lanka	Sinhala English Tamil	Sinhalese Latin Tamil	SLS 1134: 1996 - -	SLSI 1134: 2001 - -
Thailand	Thai	Thai	TIS 620: 1986	TIS 620: 1990
Vietnam	Vietnam	Latin + ext.	TCVN 5712: 1993	TCVN 5712: 1993

Source: Gordon (2005), Mikami (2002), Breton (1997) and Comrie (1990).

Table 6: Major Languages, Scripts and Character Code Standards in East/South Asia

Rohra (2005) of Saora Inc. has produced a report while doing a study to collect language corpora of Indian languages. Based on this study, it was found that user defined character encodings are the most popular followed by user community standard. For example, from more than 49 Tamil web sites visited, 15 different character sets are found. The next preferences would be the Indian Script Code for Information Interchange (ISCII) and at the bottom is a very surprising scenario for Unicode that was found to be the poorest utilized in India. Yet, based on our study by crawling India ccTLD, the top encoding is ISO-8859-1 which contributed 60%, followed by 13% for Unicode (UTF-8) and user defined encoding only represents 3%. Our large coverage has produced less limitation not only concentrating on Indian languages per se, but also other languages used in India cyberspace. This would probably be the main reason why our finding is not matching with Rohra's.

Our most recent study has disclosed that penetration of UTF-8 is limited to only 8.35% of whole web pages under Asian ccTLDs. The top three ccTLDs are Tajikistan, Vietnam and Nepal with 92.75%, 72.58% and 70.33% respectively. The least ccTLDs are Uzbekistan, Turkmenistan and Syria with very minimal 0.00% recorded. Although migration speed is expected to be high, we need to monitor carefully the process.

6 Regional Agenda

In this section, we will discuss several regional initiatives taken so far to bridge the language digital divide, with special focus on the South East Asia Regional Computer Confederation (SEARCC) region.

6.1 AFSIT/AFIT Initiatives

The Asian Forum for Standardization of Information Technology (AFSIT) was one of the early efforts to bridge the language divide and handle multilingualism. The Forum was launched in 1987 through the leadership of the Japanese Industrial Science and Technology Agency (AIST), and implemented by the Centre for International Cooperation for Computerization (CICC) with the participation of nine country representatives from the region. They were from China, Korea, India, Indonesia, Japan, Malaysia, Philippines, Singapore and Thailand. At that time, only four countries, namely Japan, China, Korea and India were actively participating in international standard development activities under the ISO/IEC JTC1 umbrella. The other countries were not able to put forth their proposals and requests on various issues of standardization to such international forums due to various reasons. The AFSIT was created to bridge this gap. The forum was organized annually and in 2002 it was renamed to Asian Forum for Information Technology (AFIT). Through these forums many country representatives receive awareness on the dire need of developing character coding standards and relevant items for their languages

On the 6th AFSIT, held in August 1992, a Special Interest Group (SIG) on Internationalization was established. The SIG meetings were held irregularly upon user needs to cater for requests such as:

- a. To clarify the concept of internationalization and to locate the target items of internationalization, namely "cultural conventions (language, scripts, culture, conventions and disposition, etc)".
- b. To distribute information on the trend and development of international activity at ISO/IEC/JTC1 to Asian countries.
- c. To locate the regional and cultural uniqueness on internationalization of IT in Asia and reflect it to the international standard.

The first SIG was held in Singapore in February 1993, and ever since it has been held almost annually for four times till November 1995. Those meetings made clear that each participating country encompasses various issues and also clarified the uniqueness and similarities of cultural conventions. In spring 1996, the results of the SIG were published in the "Data Book of Cultural Convention in Asian Countries (Sato, 1996)", and were distributed to concerned countries and parties including AFSIT and SEARCC.

6.2 SEARCC Initiatives

Under the SEARCC umbrella, several activities were also implemented. The Special Regional Interest Group on Multilingual Computing (SRIG-MLC) is one of such activities whose mission was to create a resource sharing mechanism among experts in the region in the field of multilingual computing technologies.

Multi-Lingual Computing Resources Web Site was part of the output of the SRIG-MLC activities. The site was a portal containing various data and information related to multilingual computing. One of the core experiments within this initiative was the Cyber Census project. Specifically, this project observes how many web pages exist in cyberspace by language, script, and character set. At this time the experimental sample data was only around a few hundred pages and the reports were not circulated widely.

7 Establishment of The Language Observatory Project

Recognizing the importance of monitoring language activities in cyberspace for various phenomena, the Language Observatory Project (LOP) was launched in 2003 to succeed the efforts conducted under SEARCC. Compared to an astronomical observatory, which observes space for astronomical phenomena, a language observatory observes language phenomena in cyberspace. Software agents in the form of soft bots are periodically sent into cyberspace by the mother Language Observatory in Japan to examine websites and identify its languages and contents in an attempt to identify language communities in various regions of cyberspace.

The Language Observatory project⁸ is planned to provide means for assessing the usage level of each language in cyberspace. More specifically, the project is expected to periodically produce a statistical profile of language, scripts, character code usage in cyberspace. Once the

observatory is fully functional, the following questions can be answered: How many different languages are found in the virtual universe? Which languages are missing in the virtual universe? How many web pages are written in any given language, say Pashto? How many web pages are written using the Tamil script? What kinds of character encoding schemes (CESS) are employed to encode a given

language, say Berber? How quickly is Unicode replacing the conventional and locally developed encoding schemes on the net? Along with such a survey, the project is expected to work on developing a proposal to overcome this situation both at a technical level and at a policy level. *Table 7* includes various major events as a summary of the LOP development.

Date	Event
2000, November	SEARCC/SRIG-MLC Terms of Reference established during The Second Face-to Face Meeting, Manila, Philippines
2001, March	First public version (Ver. 1.0) of the Multi-Lingual Computing Resources Web Site (http://mlcr.nagaokaut.ac.jp/mikami)
2001, November	The Cyber Census project was first openly discussed at SEARCC/SRIG-MLC meeting, Auckland, New Zealand
2002, March	A preliminary Cyber Census Experiment by Y. Mikami, I. Suzuki, Y. Chubachi , V. Narayanan and D. Rao
2002, August	A pass-breaking technique for language property identification, “Shift-Codon-Matching” published in ACM/TALIP Journal
2003, September	The Language Observatory Project was selected as one of JST sponsored program
2004, February	The First Language Observatory Work Shop (FLOWS) in conjunction with the 5th International Mother Language Day at Nagaoka University of Technology (NUT), Nagaoka, Japan. LOP was officiated by Mr Paul Hector from UNESCO.
2004, June	First crawling for The Organization of the Islamic Conference (OIC) ccTLD by UbiCrawler
2005, February	The Second Language Observatory Work Shop (LOWS2) in conjunction with the 6th International Mother Language Day hosted by UNESCO, at Tokyo University of Foreign Studies (TUFS), Tokyo, Japan

Table 7: Chronological table of events related to the birth of LOP

7.1 Project Alliance

Currently, several groups of experts are collaborating on the world language observatory. Founding organizations include: Nagaoka University of Technology, Japan; Tokyo University of Foreign Studies, Japan; Keio University, Japan; Universiti Teknologi Malaysia, Malaysia; Miskolc University, Hungary; Technology Development of Indian Languages project under Indian Ministry of Information Technology and Communication Research Laboratory, Thailand. The project is funded by Japan Science and Technology (JST) Agency under RISTEX⁹ program. UNESCO has given an official support to the project since its inception. Major technical components of the Language Observatory are basically powerful crawler technology and language property identification technology (Suzuki et al, 2002). As for crawler technology, the UbiCrawler (Boldi et al, 2004), a scalable, fully distributed web crawler developed by the joint efforts of the Dipartimento di Scienze dell’Informazione of the Università degli Studi di Milano and the Istituto di Informatica e Telematica of the Italian National Council of Research, is working as a powerful data collecting engine for the language observatory. Brief descriptions of the joint efforts of LOP and UbiCrawler team can be found in (Mikami et al, 2005).

8 Conclusion

In this paper, we stress the importance of monitoring the behaviour and activities of world languages in cyberspace. The information collected from such a study has implications on multilingual ICT education such as customized ubiquitous learning. By having a monitoring body such as that performed by the Language Observatory Project, to look at the development of languages through for an example, its encoding system, a smarter method to understand the language scenario can be realized. Through these efforts, the LOP consortium hope to make the world more aware of its living and *dying* languages in the cyberspace. Steps to assist endangered languages can then be made before its extinction. For this effort to bear fruits, the observatory is also designed to be the focal point for human capital development as well as serves to accumulate various language resources. This also includes the efforts to examine language quality based on its usage in the cyberspace. As a whole, these digital resources accumulated through research and development as well as through other means will be the bridge to lessen the digital divide. They will assist developing countries and communities to have the ability and capacity to get their indigenous language into cyberspace and hence preserves a national heritage from extinction. This capability is hoped to activate a more rigorous effort to produce more

ICT learning programs and objects in the developer mother tongue language. The LOP is also not a closed network grouping and members of SEARCC are most welcomed to participate in its activities. Having started from a SEARCC activity, it is most natural for the LOP to continue under the support and participation of SEARCC members.

9 References

- APDIP (2003). Sri-Lanka-Country Report. *Asian Forum on Information and Communication Technology Policies and e-Strategies*. 20-22 October 2003, Kuala Lumpur: Malaysia.
- Boldi, P., Codenotti, B., Santini, M. and Vigna, S. (2004) UbiCrawler: A scalable fully distributed web crawler. *Software: Practice & Experience*, 34(8):711-726.
- Breton, R. (1997). *Atlas of the Languages and Ethnic Communities of South Asia*. New Delhi: Sage Publications.
- Comrie, B. (1990). *The Major Languages of East and South-east Asia*. London: Routledge.
- Goldberg, L. (2002). Our technology future. *Education Week* found in <http://www.edweek.org/ew/newstory.cfm?slug=27goldberg.h21> Accessed on 19/04/2005.
- Gordon, G. R. Jr., ed. (2005). *Ethnologue Languages of the World*. 15th ed. Dallas: SIL International.
- Mikami, Y. (2002). *A History of Character Codes in Asia*. 1st ed. Tokyo: Kyoritsu.
- Mikami, Y., et al. (2005). The Language Observatory Project (LOP). In *Special interest tracks and posters Proceedings of the 14th international conference on World Wide Web*, May 10-14, 2005, Japan. ACM Press, 990-991.
- Narayanan V. (2004). *Indian Languages Unicode and Character Encoding. April 2004*. India Visit Report for Language Observatory Project (LOP). Nagaoka University of Technology: Japan.
- O'Neill, E. T, Lavoie, B. F, Bennet, R (2003). *Trends in the Evolution of the Public Web 1998-2002*. D-lib Magazine. April 2003: Volume 9 Number 4.
- Ravichandran, M (2005). E-Learning Initiatives in India. *Proc. International Conference on E-Commerce 2005*, Subang USJ, Malaysia, 1: 286-290, UUM.
- Rohra, A., et al. (2005). Collecting Language Corpora: Indian Languages. *The Second Language Observatory Work Shop Proceedings* 21-25 Feb. 2005. Tokyo University of Foreign Studies: Japan.
- Sato, K. T. (1996). *Data Book of Cultural Convention in Asian Countries Version 1.1*. Center of the International Cooperation for Computerization (CICC): Japan.
- Suzuki, I., Mikami, Y., Ohsato, A., Chubachi, Y. (2002). A language and character set determination method based on N-gram statistics, *ACM Transactions on Asian Language Information Processing*, 1(3): 270-279.

¹ http://www.ethnologue.com/ethno_docs/contents.asp
Accessed on 20/04/2005

² <http://www.unhchr.ch/udhr/navigate/alpha.htm>
Accessed on 15/04/2005

³ <http://www.gltreach.com/globstats/> Accessed on 28/12/2004

⁴ UNESCO (2003). *Education in a multilingual world*. Paris: Unesco, (ED-2003/WS/2).
<http://unesdoc.unesco.org/images/0012/001297/129728e.pdf>

⁵ http://www.nectec.or.th/pub/book/e_index.html

⁶ <http://www.iana.org/assignments/character-sets>

⁷ <http://www.itscj.ipsj.or.jp/ISO-IR/>

⁸ <http://www.language-observatory.org>

⁹ http://www.ristex.jp/english/top_e.html

Acknowledgement

The study was made possible by the financial support of the Japan Science and Technology Agency (JST) under the RISTEX program. We also thank UNESCO for giving official support to the project since its inception.