The International Journal on Advances in ICT for Emerging Regions 2008 01 (01) : 12 - 23

An Analysis of Asian Language Web Pages

S. T. Nandasara[1*], Shigeaki Kodama[2], Chew Yew Choong[3], Rizza Caminero[4], Ahmed Tarcan[5], Hammam Riza[6], Robin Lee Nagano[7], Yoshiki Mikami[8]

[1] University of Colombo School of Computing, Colombo, Sri Lanka

[2, 3, 4, 8] Nagaoka University of Technology, Nagaoka, Niigata, Japan

[5] Dicle University, Diyarbakir, 21280, Turkey

[6] IPTEKnet, BPPT, Indonesia

[7] Miskolc University, Miskolc, Hungary

stn@ucsc.cmb.ac.lk, kodamas@kjs.nagaokaut.ac.jp, yewchoong@hotmail.com, rhyze1018@yahoo.com, tarcan@dicle.edu.tr, hammam@iptek.net.id, nagano.robin@chello.hu, mikami@kjs.nagaokaut.ac.jp

Revised: 24 September 2008; Accepted: 24 July 2008

Abstract: This paper gives an overview and an evaluation of Web pages of Asian languages on the Web, in particular of those languages that have not been focused on so far. The authors have collected over 100 million Asian Web pages downloaded from 42 Asian country domains, identified the languages based on N-gram statistics and analyzed their language properties. Primarily the number of pages written in each language measures the presence of a language. The survey reveals that the digital language divide exists at a serious level in the region. The state of multilingualism and the dominating presence of cross-border languages, English in particular, are analyzed. The paper sheds light on script and encoding issues of Asian language texts on the Web. In order to promote language resource collection and sharing, authors have a vision of creating an observation-collection instrument for Asian language resources on the Web. The results of the survey show the feasibility of this vision, and provide us with a better idea of the steps needed to realize that vision.

Keywords: Asian languages, Data Mining, Web Statistics, Language Identification, Standards, Multilingualism, Encoding, Web as Corpus, Digital Language Divide.

INTRODUCTION

Since the early days of Web development, various attempts have been made to grasp the language distribution of the Web. An estimate of language distribution in terms of Internet users' languages has been regularly reported by a marketing research group [1]. Estimates of the distribution of the Web documents are compiled by various groups, each with a different scope and focus. The work of Alis Technologies and the Internet Society [2] is among the earliest. Network and Development Foundation (FUNREDES) compiles a regular report focused on the Romance language group [3], and Online Computer Library Center's (OCLC) Web Characterization Project [4] covers large number of European languages. Most of these surveys have evolved along with the multilingual search engines like Inktomi, Yahoo, Google, Alltheweb, etc. The language-specific search capability of the search engines has provided a means of surveying for researchers. Although these surveys have given us fairly good pictures about European language presence on the Web, far less attention has been paid to Asian languages, among them "less computerized languages" in particular.

This ignorance may arise partly from the technical difficulties of language identification of Asian languages and partly from "commercial value" of Asian languages that has been low. With the exceptions of Chinese, Japanese, Korean, Thai, Malay, Turkish, Arabic, and Hebrew, nothing is known about the extent of the presence of Asian languages on the Web. We felt a strong need to implement an independent survey instrument to observe the activity level of those languages. The UNESCO report, presented to the Tunis phase of the World Summit on the Information Society, *"Measuring Language Diversity on the Internet"* [5], shares exactly the same concerns as we do.

In response to this, the Language Observatory (LO) project was launched in 2003 under the sponsorship of the Japan Science and Technology Agency (JST) and has been implemented in collaboration with several international partners who have common interests with us [6]. After a few years of development work, the LO team has trained our own language identification engine to cover more than three hundred languages of the world, and has acquired the capability to collect terabyte size Web documents from the Internet. The paper is based on the preliminary survey results of this project.

In addition, we have begun a sister project, the Asian Language Resource Network project by the sponsorship of Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) from 2005. We find a synergy between those two projects: the observation instrument for Asian languages can work as a language resource collection instrument as well. We have a vision of integrating the two projects as an observation-collection instrument for Asian language resources on the Web.

* *corresponding author*

OBJECTIVES

The objectives of this paper are firstly to give an overview for Asian languages on the Web, in particular for those languages that have been ignored up to now. Through this study, we have tried to spotlight the presence of Asian languages. Here the presence of a language is measured primarily by the number of pages written in each language and is supplemented by additional indicators like pages-per-capita to give an indication of the relative intensity of Web authorship. In terms of language coverage, we discovered 55 Asian languages. Chinese, Japanese and Korean are excluded from the analysis because the presence of these languages can be relatively easily measured by using existing commercial search engines.

Secondly, the paper tries to describe the state of multilingualism in Asian country domains. The state of multilingualism can be defined at various levels, from a personal or document level to a societal level. In this study, we show a multiple language presence in each country domain. To give an overview of cross-border languages is a part of these efforts.

Thirdly, the paper tries to shed light on script and encoding issues of Asian languages. The paper tries to answer questions like; to what extent is UCS/ Unicode employed for Asian languages? What scripts are actually used to represent a specific language? To what extent are locally developed encodings used? Most European languages are written in only one script, Latin, Cyrillic or Greek. Some Asian languages, however, are written in a variety of scripts. This is most notable in the central Asian region, where the same language can be presented in Cyrillic, Arabic and Latin. In addition, each script is presented in various encoding schemes. While UCS/Unicode is expected to play a pivotal role in the promotion of multilingual document processing on the Web, its actual implementation on the Web seems still very much limited. Instead, various local legacy encoding schemes are employed. In order to promote language resource collection and development, due attention should be paid to script and encoding variety issues, particularly in this case, where it leads to a chaotic situation of encoding as observed in Asian language documents on the Web.

And lastly, the section three discusses the data collection using *UbiCrawler*, language identification process including creating training data sets and analytical methodologies. The Asian language presence on the Web is discussed in "Asian Language Presence on the Web" section. The state of multilingualism and the presence of cross-border languages are discussed in "Multilingualism in the Asian Web" section and script and encoding issues are discussed in " Script and Encoding Issues" section . Finally discussions on future research areas and conclusion are given in "Descussion" and "Conclusion" sections respectively.

METHODOLOGY

Web Pages Collected

We use a Web crawler that works by downloading Web pages from the Internet. While downloading, it traces links within pages and recursively crawls to gather those newly discovered pages. The collection of downloaded Web pages is then passed to the language identification engine and the language properties of the pages are identified.[7].

The latest Asia crawl (excluding China, Japan and Korea) focused on Web pages in 42 country domains in Asia. The crawl was begun from a seed file containing 13,286 URLs . The list of ccTLDs (country code Top Level Domains) contains ae, af, az, bd, bh, bn, bt, cy, id, il, in, iq, ir, jo, kg, kh, kw, kz, la, lb, lk, mm, mn, mv, my, np, om, ph, pk, ps, qa, sa, sg, sy, th, tj, tm, tp, tr, uz, vn and ye. Web pages outside of these ccTLDs were not crawled. The crawl was performed using a decentralized, parallel crawler called UbiCrawler [8][9]. The crawler is configured to stop tracing further links at a depth of 8 and to download a maximum of 50,000 pages per site. The crawler waits 30 seconds for http header responds before giving up.

The Asia crawl started from 5th July 2006 at 11:00hrs and ended on 19th July 2006 at 19:03hrs without any problem. We downloaded 107,141,679 Web pages in total, 652,710,237,381 bytes in size.

UbiCrawler supports the Robot Exclusion standard and we fully respect it at all Web sites. The crawler is configured to check and analyze *Robots.txt* on every new Web site. If a Web site indicates Web robots are not welcome, our crawler will not download that Web site. The latest Asia crawl discovered 45,348 Robots.txt files.

Further, Web sites and their contents change over time. Most search engines have accumulated their databases to have longer (in time) coverage. This means that in the database, there might be many obsolete Web sites and pages. Because the pages downloaded during one short period of time in our study accurately reflect the "current" status of Web sites.

Lastly, while search engines generally cache all types of files, we only crawl for html and text files, both static and dynamic. Although there are many documents available in PDF format, we excluded PDF files because of technical difficulties in handling PDF for Language Identification Module (LIM).

Language Identification Process

The Language Identification Module (LIM) developed for the Language Observatory Project (LOP) [10] can simultaneously detect the triplet of Language, Script and Encoding (LSE) scheme (LSE is used below for this triplet) for each document. The identification is based on the n-gram statistics of documents. A natural language model, which assumes that the probability of the next word depends on the previous few words

is generally known as an N-gram model, and a series of N characters (or N bytes) can be referred to as an "N – gram" as well. The advantages of the n-gram approach are that it does not require a special dictionary or word frequency list for each language, and it can detect encoding scheme.

LIM consists of two components. First, the training component accumulates sets of shift-codons from the training data. The term "shift-codon" is derived from the genetic term "codon", a sequence of three nucleotides. Shift-codons are, as the naming implies, three byte strings extracted from the first position, the second position, (n-2)-th position of a training data (n is the length of the training data). The set of shift-codons thus created are stored with the LSE tags into the reference database. The source of training data is translations of the Universal Declaration of Human Rights (UDHR) provided by the United Nation's Office of Higher Commissioner for Human Rights.

The second component, the identification component, produces shift-codons of the target data and then compares them with all sets of shift-codons stored in the reference database. After comparison, the component calculates the matching ratio of the shift-codons of the target text to those of the training text (the number of matched codons of the target document divided by the total number of codons). Then the component returns the LSE that shows the highest matching ratio as a result. The component returns "Below Threshold" when the highest matching ratio is below a given threshold, and returns "No Match" when no single codon of the target document matches with those of stored reference data. The component returns "Short" or "Empty" when the byte length of the target document is not enough to be identified or no content is found on the target document after removing HTML tags.

There are two data sets used in the language identifier (LI). First, it is the data set that we use to train the LI; we called it the Training corpus (TC). The second data set is the Validation corpus (VC), which contains 500 multilingual Web pages that manually checked by users to confirm their actual language, script and encoding (LSE). The purpose of this corpus is to ensure the accuracy of the LI. Since we already know the correct LSE of the VC, every time we made changes to the LI, we can perform an experiment against the VC and find out how the changes affect the accuracy rate.

The language identification engine LI has been trained in more than 200 languages of the world (345 in terms of LSEs) at the time of this survey. Among them, 62 languages are spoken in Asia and total of 98 different encodings for Asian language scripts have been trained. Missing Asian languages from the UDHR listing are Zhuang, Yi, Hmong (including its various dialects), Shan, Karen, Oriya, Divehi, Dzongkha (Bhutanese), etc.

Languages selected here are official or nationally recognized languages in respective Asian countries. Training data sets are based on the Universal Declaration of Human Rights document, which has been converted into each language and into commonly used encoding schemes including UTF-8. Table 1 is the complete list of the Asian languages targeted in this survey, classified by language family. Additional information for the languages is also listed, viz: the script(s) for the language and the encodings we trained LIM over.

ASIAN LANGUAGE PRESENCE ON THE WEB

Introduction to Asian Languages

We can list several language families on the Asian continent: Austroasiatic, Austronesian, Dravidian, Indo-Iranian, Mongolian, Semitic, Sino-Tibetan, Thai-Kadai, Turkic, and Tungus. Some of these language families are not firmly established and could be regrouped into the larger language groups or could be divided into smaller sub-groups. For example, the Turkic, Mongolian, and Tungus language families can be regrouped into larger language family Altaic, and the Indo-Iranian language family can be divided into the Indo-Aryan, Iranian, and Kafiri. There are some isolated languages around the Asian continent, e.g. Korean, Japanese, Ainu, and Burushaski. Some European languages – English, Russian, French, and Portuguese – are also used in the region as official languages, and from the mixture of an indigenous language and an introduced language, pidgins or creoles have emerged.

Among those language families, Sino-Tibetan has the largest number of speakers, estimated at 1.2 billion. Next comes Indo-Iranian, with at least 700 million speakers in India, and more than 200 million people in Pakistan, Bangladesh, Iran and other South and MiddleEast Asian countries. Malay in the Austronesian language family has around 250 million speakers in Indonesia, Malaysia, Brunei, Singapore, the southern Philippines, and Thailand. Tamil, a Dravidian family, has about 200 million speakers in India. Semitic includes a language of many speakers, that is, Arabic, the number of which is estimated to be about 200 million. Other language families have a relatively small number of speakers. Among the isolated languages, Japanese has the largest number of speakers with about 125 million and Korean follows with about 75 million. When we describe the Asian languages, we cannot avoid mentioning the diversity of scripts they use. Contrasted with Western Europe, the diversity is outstanding. In Southeast and South Asian countries, many scripts that come from the Brahmi script are used, and in the East and Near East Asian countries, Hanzi script and some other indigenous scripts are used. Latin, Arabic and Cyrillic script are also used with some additional letters and diacritical marks.

Web Presence by Country

In Figure 1, the colouring of map is based on the number of Web pages per 1000 population, as this is the reflection of the degree of presence of a country on the

Table 1: List of Language/Script/Encoding[1] trained, grouped by language family

[Austronesian]	[Indo-Iranian]	[Dravidian]
Achehnese/Latin/Latin1	Assamese/Bengali/UTF-8	Kannada/Kannada/UTF-8
Balinese/Latin/Latin1	Balochi/Arabic/UTF-8	Tamil/Tamil/UTF-8
Bikol/Bicolano/Latin/Latin1	Bengali/Bengali/UTF-8	Tamil/Tamil/Vikata
Bugisnese/Latin/Latin1	Bhojpuri/Devanagari/Agra	Tamil/Tamil/Shree
Cebuano/Latin/Latin1	Dari/Arabic/UTF-8	Tamil/Tamil/Kumudam
Filipino/Latin/Latin1	Farsi/Persian/Arabic/UTF-8	Tamil/Tamil/Amudham
Hiligaynon/Latin/Latin1	Gujarati/Gujarati/UTF-8	Telugu/Telugu/UTF-8
Indonesian/Latin/Latin1	Hindi/Devanagari/UTF-8	Telugu/Telugu/TLW
Javanese/Latin/Latin1	Hindi/Devanagari/Naidunia	Telugu/Telugu/Shree
Kapampangan/Latin/Latin1	Hindi/Devanagari/Arjun	
Iloko/Latin/Latin1	Hindi/Devanagari/Shusha	**[Semitic]**
Madurese/Latin/Latin1	Hindi/Devanagari/Shivaji	Arabic/Arabic/UTF-8
Malay/Latin/Latin1	Hindi/Devanagari/Sanskrit	Arabic/Arabic/Arabic
Minangkabau/Latin/Latin1	Hindi/Devanagari/Kiran	Hebrew/Hebrew/UTF-8
Sundanese/Latin/Latin1	Hindi/Devanagari/Hungama	Hebrew/Hebrew/Hebrew
Tetun/Latin/Latin1	Hindi/Devanagari/Shree	
Waray/Latin/Latin1	Hindi/Devanagari/KrutiDev	**[Turcic]**
	Kashimiri/Devanagari/UTF-8	Abkhaz/Latin/UTF-8
[Austro-Asiatic]	Kurdish/Latin/UTF-8	Abkhaz/Cyrillic/8859-5
Hmong/Latin/Latin1	Magahi/Devanagari/UTF-8	Abkhaz/Cyrillic/Abkh
Khmer/Khmer/UTF-8	Magahi/Devanagari/Agra	Azeri /Latin/Az.Times
Vietnamese/Latin/UTF-8	Marathi/Devanagari/KrutiDev	Azeri /Cyrillic/Az.Times
Vietnamese/Latin/TCVN	Marathi/Devanagari/Shivaji	Kazakh/Cyrillic/8859-5
Vietnamese/Latin/VIQR	Marathi/Devanagari/Kiran	Kazakh/Arabic/UTF-8
Vietnamese/Latin/VPS	Marathi/Devanagari/Shree	Tatar/Latin/Latin1
	Nepali/Devanagari/UTF-8	Turkish/Latin/UTF-8
[Sino-Tibetan]	Osetin/Arabic/UTF-8	Turkish/Latin/Turkish
Burmese/Burmese/UTF-8	Osetin/Cyrillic/UTF-8	Uighur/Latin/UTF-8
Chinese/Hanzi/GB2312	Pashtu/Arabic/UTF-8	Uighur/Latin/Latin1
Chinese/Hanzi/UTF-8	Punjabi/Arabic/UTF-8	Uzbek/Latin/Latin1
Hani/Latin/Latin	Sanskrit/Devanagari/UTF-8	
Tamang/Devanagari/UTF-8	Saraiki /Arabic/UTF-8	**[Thai-Kidai]**
Tibetan/Tibetan/UTF-8	Sinhala/Sinhala/UTF-8	Lao/Lao/UTF-8
	Sinhala/Sinhala/Kaputa	Thai/Thai/TIS620
[Mongolian]	Sinhala/Sinhala/Metta	Thai/Thai/UTF-8
Mongolian/Cyrillic/UTF-8	Tajiki/Arabic/UTF-8	Zhuang/Latin/Latin1
Mongolian/Cyrillic/8859-5	Urdu/Arabic/UTF-8	

[1] Local proprietary encodings are shown in this table by names of font.

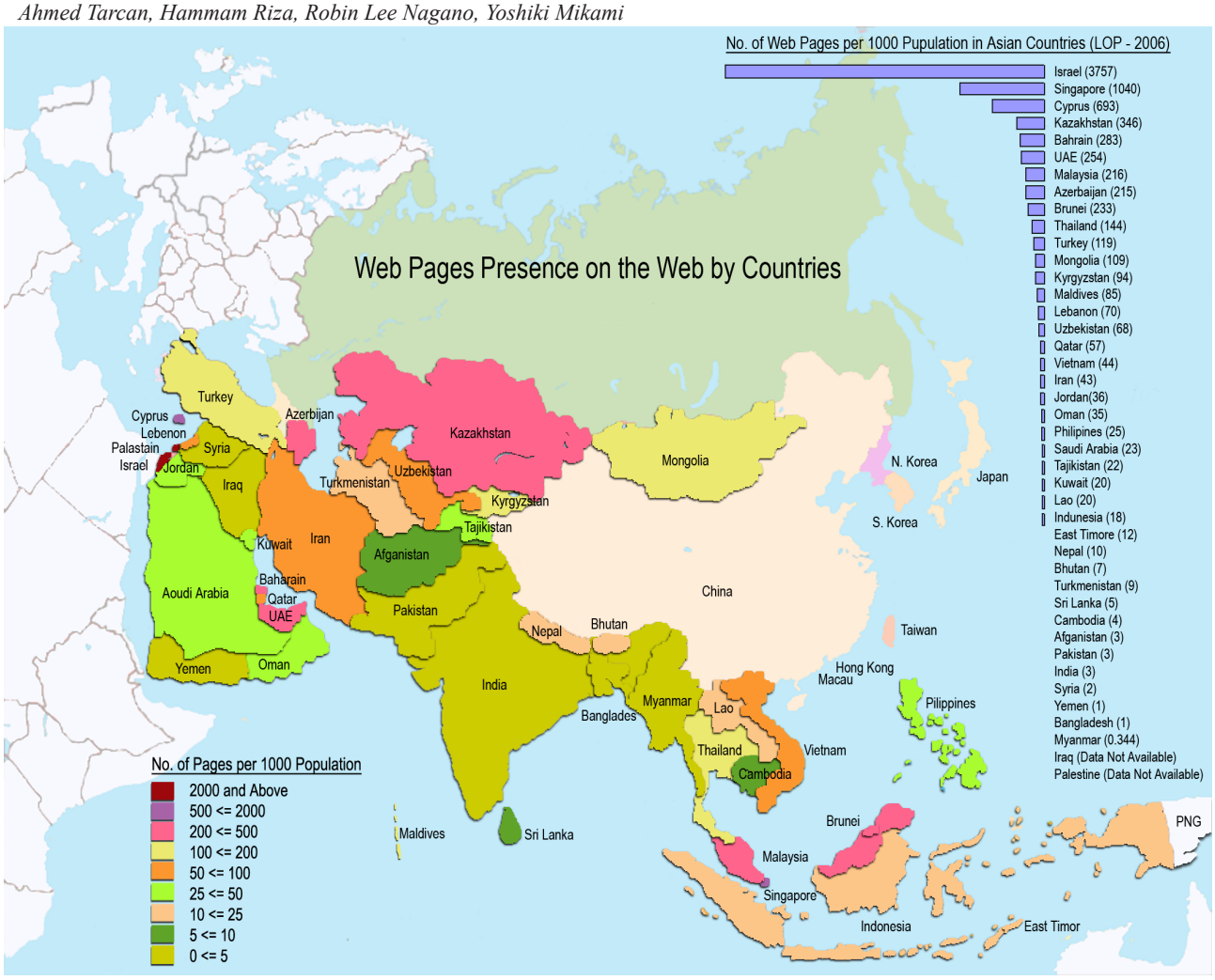

Figure 1*:* Presence of Web Pages by Country in the Asian Region

Web. This shows that Israel is the highest (3757 pages per 1000 population) in the rank and Singapore and Cyprus follow, respectively. The population data was obtained from the CIA World Factbook (estimates as of July 2006).

Figure 1 shows that Kazakhstan and Azerbaijan respectively have the highest Web page size per 1000 population among Central Asian countries. Figure 1 also shows that Cambodia, Afghanistan, Pakistan, India, Syria, Yemen, Bangladesh, and the last, Myanmar, have the least number of pages presence on the Web (between 5 (4.54%) to 0 (0.35%) pages per 1000 population). It is worth noting that Myanmar, the neighboring country to Thailand, has the least (0.35%) among all the Asian countries.

The presence on the Web of each Asian country is given at ccTLD level in Table 2. In Table 2, ranking is based on the percentage of Web presence against the total Web pages in the region. This shows that Israel (28.88%), Thailand (11.72%) and Turkey (10.61%) with a higher number of language presence on the internet at ccTLD level. Table 2 was tabulated using the Number of Web Pages collected by the crawler engine.

Table 2: Percentage of Web Pages on the Internet at ccTLD Level

ccTLD	% of Web Pages	ccTLD	% of Web Pages	ccTLD	% of Web Pages
il	28.88	ae	0.87	lk	0.13
th	11.72	kg	0.69	bn	0.09
tr	10.61	pk	0.69	ps	0.08
my	6.41	cy	0.59	tm	0.08
kz	6.01	mn	0.37	kh	0.06
sg	5.39	np	0.37	kw	0.06
id	5.36	lb	0.32	qa	0.05
vn	4.19	jo	0.27	sy	0.05
in	3.98	bh	0.23	bt	0.04
ir	3.75	tj	0.22	mv	0.03
ph	2.55	bd	0.19	ye	0.03
uz	2.13	la	0.14	mm	0.02
az	2.10	om	0.14	tp	0.01
sa	0.98	af	0.13	iq	0.00

Web Presence by Language

Fourth column of Table 3 shows the total number of Web pages identified by the survey. The data shown in the third column of the table is the speaker population of that language with statistics taken from the UDHR

Web site. In principle, all Asian languages listed in first column in Table 3 are considered as local languages. The ranking is based on the number of pages. Table 3 shows that Hebrew, Thai, Turkish, Vietnamese, Arabic, Tatar, Farsi, Javanese, Indonesian, Malay, Sudanese, Hindi, Dari, Uzbek and Mongolian have a relatively high presence on the Web. The highest number is for Hebrew, and the second highest for Thai. The fifth column gives the number of pages per 1000 speakers of each language. An almost identical ranking is observed in both the number of pages and the pages per population.

A high degree of "divide" in terms of usage level of languages can be observed among the Asian languages. The number of Hebrew pages per 1000 speakers is 28 times higher than that of the Malay language (ranked tenth in Table 3), 300 times higher than Pushtu (ranked 20th), and 3,000 times higher than Gujarati (ranked 50th). The speakers' population of languages is said to follow Zipf's Law - the n-th ranked language speaker is one n-th of the population of the top ranked language. But if we measure the size of a language by the number of pages written in the respective language, the relative size of the 1st, 10th, 20th and 50th ranked language in Table 3 becomes a series of 1, 0.036, 0.0035, 0.0001. Our observation suggests that the number of Web pages written in each language follows a progressive power law curve. The situation evidenced here can be well described as a "Digital Language Divide".

MULTILINGUALISM IN THE ASIAN WEB

Multilingualism by Country Domain

The most recent version of Ethnologue [11] lists close to seven thousand languages around the world. More than 2600 of them are spoken in the Asian region. This indicates that a large scale linguistic diversity is observable in Asia. Among the 2600s', only around 51 languages are recognized by Asian governments as official or national language(s) of the country and other languages have been recognized as languages for home use.

Through the survey, a rich diversity of written pages was found in the country with the richest diversity of languages in the region, i.e. Indonesia. It is interesting to note that there are a significantly larger number of pages in Javanese compared to either Indonesian or Malay. The major language found in Indonesia, Malaysia, Brunei, Singapore, Southern Thailand and Phillipines can be categorized into a single root Malay language spoken in different dialects. This surprising result shows two things: Javanese has a dominating Web presence in Indonesia. The lesser Sundanese, Madurese, Achehnese and Buginese languages are found to be of great importance to Indonesia's local language diversity on the Internet (see Table 3).

Table 3: Number of Web Pages Collected from Asian ccTLDs, by Language

Language	Script	Speaker population	Total number of pages	No. of pages per 1000 speakers
Hebrew	Hebrew	4,612,000	11,957,314	18.08
Thai	Thai	21,000,000	7,752,785	11.72
Turkish	Latin	59,000,000	3,959,328	5.99
Vietnamese	Latin	66,897,000	2,006,469	3.03
Arabic	Arabic	280,000,000	1,671,122	2.53
Tatar	Latin	7,000,000	1,575,442	2.38
Farsi	Latin	33,000,000	1,293,880	1.96
Javanese	Latin	75,000,000	1,267,981	1.92
Indonesian	Latin	140,000,000	866,238	1.31
Malay	Latin	17,600,000	432,784	0.65
Sundanese	Latin	27,000,000	217,298	0.33
Hindi & others	Devanagari	182,000,000	119,948	0.18
Dari	Arabic	7,000,000	107,963	0.16
Uzbek	Latin	18,386,000	57,212	0.09
Mongolian	Cyrillic	2,330,000	51,140	0.08
Kazakh	Arabic	8,000,000	48,652	0.07
Madurese	Latin	10,000,000	47,246	0.07
Uighur	Latin	7,464,000	46,399	0.07
Kashmiri	Arabic	4,381,000	41,876	0.06
Pushtu	Arabic	9,585,000	41,479	0.06
Balochi	Arabic	1,735,000	36,497	0.06

Language	Script	Speaker population	Total number of pages	No. of pages per 1000 speakers
Turkmen	Latin	5,397,500	32,156	0.05
Minangkabau	Latin	6,500,000	20,766	0.03
Bikol	Latin	4,000,000	18,509	0.03
Kyrgyz	Arabic	2,631,420	15,606	0.02
Balinese	Latin	3,800,000	14,584	0.02
Punjabi	Arabic	25,700,000	14,544	0.02
Sindhi	Arabic	19,675,000	12,945	0.02
Achehnese	Latin	3,000,000	11,102	0.02
Sinhala	Sinhala	13,218,000	10,770	0.02
Kapampangan	Latin	2,000,000	10,094	0.02
Iloko	Latin	8,000,000	9,180	0.01
Bengali & Assamese	Bengali	196,000,000	8,590	0.01
Filipino	Latin	14,850,000	5,511	0.01
Waray	Latin	3,000,000	5,426	0.01
Bugisnese	Latin	3,500,000	3,533	0.00
Burmese	Burmese	31,000,000	3,285	0.00
Kurdish	Latin	20,000,000	3,135	0.00
Tajiki	Arabic	4,380,000	2,430	0.00
Azeri	Cyrillic / Latin	13,869,000	3,767	0.00
Tamil	Tamil	62,000,000	2,025	0.00
Hiligaynon	Latin	7,000,000	1,935	0.00
Dhivehi	Thaana	250,000	1,858	0.00
Bhojpuri	Devanagari	25,000,000	1,756	0.00
Tibetan	Tibetan	1,254,000	1,454	0.00
Cebuano	Latin	15,230,000	1,107	0.00
Telugu	Telugu	73,000,000	1,072	0.00
Saraiki	Arabic	15,020,000	1,036	0.00
Lao	Lao	4,000,000	799	0.00
Gujarati	Gujarati	44,000,000	765	0.00
Pashto	Arabic	9,585,000	259	0.00
Kannada	Kannada	33,663,000	164	0.00
Urdu	Arabic	54,000,000	70	0.00
Khmer	Khmer	7,063,200	65	0.00
Hani	Latin	747,000	63	0.00
Asian Languages total (A)			33,838,551	(51.2%)
Other Languages total (B)			32,293,912	(48.8%)
Identified pages total (A + B)			66,132,463	(61.7%)
Unidentified pages total (C)			41,009,216	(38.3%)
Matching ratio below threshold [1]			5,701,765	(5.3%)
Empty pages			273,187	(0.3%)
No matching pages			9,386	(0.0%)
Duplicated pages [2]			35,024,878	(32.7%)
Total downloaded Pages (A + B + C)			107,141,679	(100%)

[1] The threshold is set at 20% in this survey; [2] Almost one-third of the pages were found to be an exact copy of another page. We excluded duplicate pages from the language identification process.

Cross-Border Languages and their Dominance

Another aspect of the multilingualism in the region is the overwhelming presence of cross-border languages on the Web. Here we define two categories of languages. The first category is "local languages", which are officially recognized language(s) and home speakers' languages of the state. The second category is "cross-border languages", such as English, French, Russian and Arabic, which are used as a language of communication among the peoples of different nations.

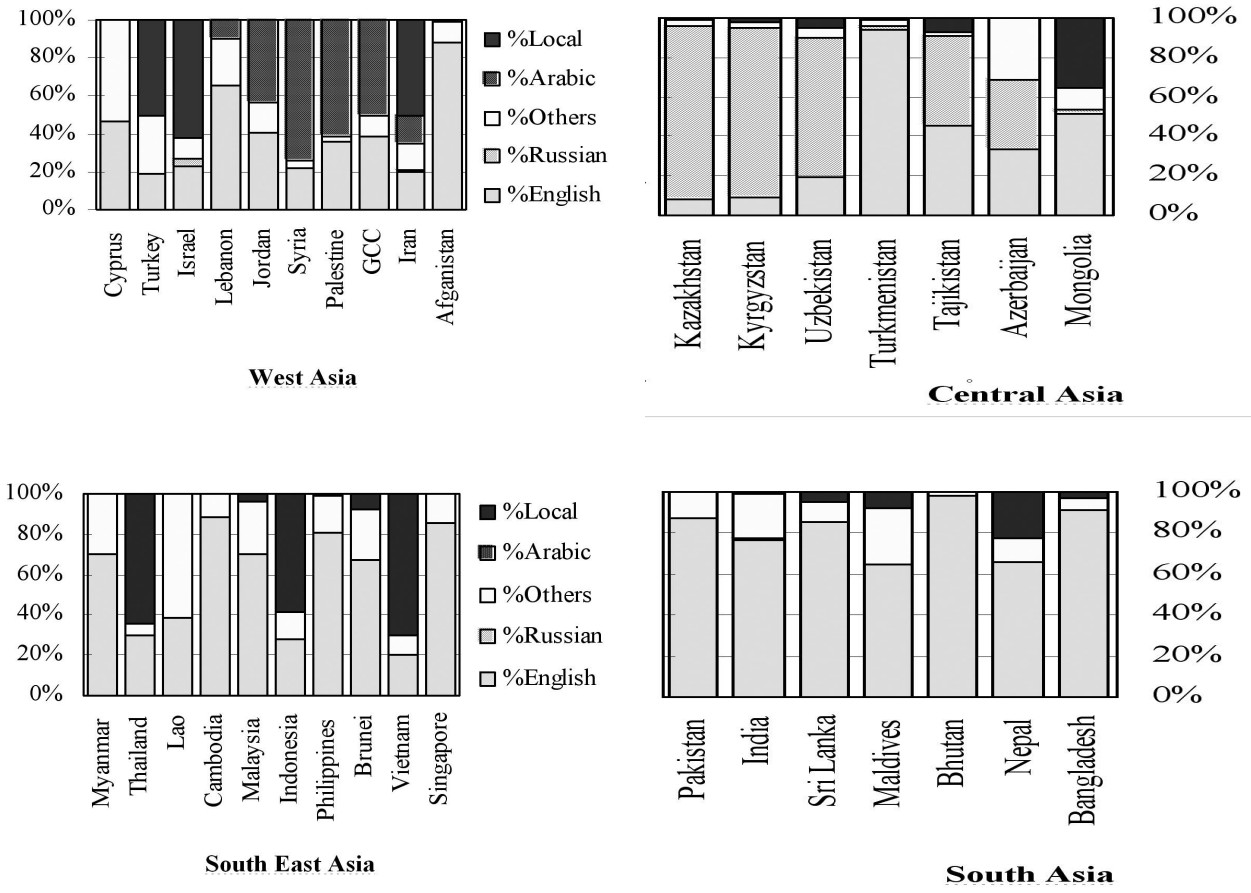

Figure 2: Cross-border languages presence in Asian countries grouped by region

GCC stands for the Gulf Coopeation Coucil, which consists of Bahrain, Kuwait, Oman, Qatar, Saudi Arabia and UAE

Arabic can be categorized in two ways. In the Middle East region, Arabic is recognized as an official language in many countries, but it also functions as an important cross-border language. So we treat Arabic in two ways depending on the context of analysis; if it is an official language, it is counted as a local language in Figure 2, and if not, then as an 'other cross-border language' [12]. Figure 2 shows the relative share of these categories of language in each country domain. Countries are grouped by sub-region. We found that each sub-region shows clear characteristics in terms of the weight and the choice of cross-border languages.

In West Asia, two cross-border languages, English and Arabic, dominate the Web. Almost 99% of Web pages are written in these two cross-border languages, except in Cyprus, Iran and Israel. Local languages show a majority in several countries, such as in Israel (62.0% in Hebrew), Turkey (50.7% in Turkish) and Iran (50.6% in Farsi, Dari, Pashtu and Balochi). If we treat Arabic as a local language, the share of local languages becomes more than half in most countries in the region. A quite unique case is Cyprus, where Greek (36.6%) plays a key role.

In South Asia, the dominance of English is outstanding. Relatively high share of local languages is found only in Nepal (22.4% in Hindi or Nepali), India (21.7% in various Indian languages), the Maldives (8.2% in Divehi) and Sri Lanka (5.1% in Sinhala).

In the Central Asia, there are two cross-border languages, English and Russian. As will be discussed in section " The Waro Alphabets in Central Asia", Russian dominates in Kazakhstan (88.8% in Russian), Kyrgyzstan (86.3% in Russian), and Uzbekistan (70.4% in Russian), while English dominates in Turkmenistan (94.4% in English). Tajikistan (45.9% in English and 44.6% in Russian) has an equal balance of the two languages. Local languages show a substantially higher share (35.2% Mongolian) only in Mongolia.

In South East Asia, the situation is rather different. The share of local languages is far higher than in other sub-regions. Among them, local languages have a major share in Vietnam (69.8% in Vietnamese), Thailand

(64.0% in Thai) and Indonesia (58.7% in various local languages including Javanese, Indonesia, Sundanese, Balinese, etc.). English dominance is also observed frequently in this sub-region.

SCRIPT AND ENCODING ISSUES

Script Diversity in Asia

Asia is especially rich in scripts. The five basic scripts: Ideographic, Brahmi, Latin, Arabic and Cyrillic grew up in the region, each largely separated by mountains, ocean or deserts. In East Asia, the influence of Chinese ideographic script (hanzi) is remarkable. In South Asia, in and around the Indian Subcontinent and in the continental part of Southeast Asia, scripts originating from Brahmi-script are influential. The islands of Southeast Asia and Australasia have mostly adopted Latin scripts (some islands in the region still use Brahmi-originating scripts such as the Balinese script, or aksara Bali). In Central Asia, historically languages were written in the Arabic script under the influence of the Ottoman Empire but later transformed into Cyrillic. Lastly in the western part of Asia, Arabic script is widely used not only by Arabic speakers but also by non-Arabic speakers.

Table 4: Number of Pages in Domains of Central Asian Republics

(a) English, Russian, and Arabic pages by country

Country	English (A)	Russian (B)	Arabic (C)	(A + B + C)
Azerbaijan	553,168	534,913	3,081	1,091,162
Kazakhstan	263,125	2,234,674	106	2,497,905
Kyrgyzstan	42,167	403,080	55	445,302
Tajikistan	48,300	45,178	27	93,505
Turkmenistan	1,398,708	5,922	4,004	1,408,634
Uzbekistan	255,782	922,188	15	1,177,985
Total	2,561,250	4,145,955	7,288	6,714,493

(b) Official Language pages by Script

Language	Latin (A)	Cyrillic (B)	Arabic (C)	(A + B + C)
Azeri	726	2,315	n/a	3,041
Kazakh	n/a	48,522	130	48,652
Kyrgyz	n/a	12,680	2,962	15,642
Tajiki	n/a	n/a	2,430	2,430
Turkmen	32,156	n/a	n/a	32,156
Uzbek	57,212	n/a	n/a	57,212
Total	90,094	63,517	5,522	159,133

"n/a" means pages are not yet found, but it does not mean non-existence of pages.

The War of Alphabets in Central Asia

As the Turkic language border extends from Europe to China, covering 12 million square kilometres, the languages are written in several scripts. In Turkey the Latin alphabet has been used since 1928. In Central Asian republics, the Cyrillic script has been in use from about the same time. In some areas of Afghanistan Arabic script is used. It is said that Turkic languages such as Uyghur and Kazakh are written even in Chinese script. Now the region is in a transition period. As an interesting example of script diversity, let us discuss this sub-region.

Since 1990 Turkey has invited thousands of students from new republics in the central Asia to Turkish universities by offering scholarships. The Turkish Education Minister Köksal Toptan, while attending a conference of the education ministers of Turkic republics and communities in Bishkek in September 1993, said, "the most important factor which will secure our unity and develop our language is a common alphabet" [13]. Azerbaijan, Turkmenistan, Uzbekistan, Tatarstan, and Gagauz have an agreement to complete the transition into Latin script before 2010.

But in reality, in symposiums and meetings between Turkic republics in Central Asia Russian is nearly the sole tool of communication between the Turkic peoples in the central Asia. "The local languages are used exclusively in indigenous film-making, scholarly publication, and in local trade and commerce" [13] Kazakhstan and Kyrgyzstan have a significant Russian population. This fact increases again the influence of Russian language and the Cyrillic script as well. China and Iran are the other important actors in this sub-region: Kyrgyzstan and Kazakhstan share borders with China, and Iran has an important influence on Azerbaijan and Turkmenistan.

Although our survey results can provide only a limited picture of this situation, they do make it clear that the choice of script used to write local languages seems influenced substantially by the script of the dominating language in the country (Table 4(a), 4(b)).

The Existence of Multiple Encodings

Indian language Web sites heavily rely on unique encodings or proprietary extensions of existing standard encodings [14]. One survey had found 24 such local encodings for Hindi alone, and 15 for Tamil, 14 for Marathi, 10 for Malayalam, and so on. The total number of these local encodings reaches well over 50 [15]. The existence of multiple local encodings is not specific only to Indian languages, but is widely seen in other languages which use non-Latin scripts or Latin script with significant extensions and/or additional diacritics. Vietnamese is a typical example of the latter.

To resolve this problem for scripts of the languages around the world, the International Organization for Standardization (ISO) and the International Electro-technical Commission (IEC) made efforts to develop a single comprehensive universal character set. The first version of "The Universal Multiple Octet Coded Character Set (UCS)" was published in 1991. Later the

work of ISO / IEC and that of the Unicode Consortium became integrated and synchronized. The most recent version of the Unicode Standard (The Unicode Consortium, 2005) assigns a unique identifier to each of 97,720 characters (including 70,207 ideographic characters defined by national and industry standards of China, Japan, Korea, Taiwan, Vietnam and Singapore). But it was expected by the Unicode Consortium, that encoding based on UCS/Unicode, whose most commonly used form is UTF-8 (UTF-Unicode Transformation Format), would be used in parallel with the above-mentioned local encodings.

Taking this plurality into account, we have tried to collect training data encoded in these local proprietary codes and in UTF-8. As shown in Table 5, we have trained our language identification engine LIM by using 9 encodings for Hindi, 4 encodings for Tamil and 2 encodings for Telugu. Also we have included 3 encodings for Vietnamese and 2 encodings for Sinhala. This is still not sufficient to match the plurality in the real world, but we believe that this is the first ever attempt to identify actual usage of local encodings in the Web space.

As a result, we found that the use of UTF-8 in the Asian region is extremely low. Table 5 shows to what extent UTF-8 is used in selected languages (for other languages, we do not have sufficient training data prepared in different encodings). The table shows that Vietnamese is found to be the highest in the penetration of UTF-8 (96.4%). It is followed by Mongolian (95.5%), Hindi and other Devanagari-based languages (78.4%), and Sinhala (44.5%). Hebrew (12.3%), Thai (2.7%), Burmese (0.7%), and Turkish (0.5%) are relatively or extremely low. These estimates of UTF-8 penetration should be considered as overestimated, because many local encodings are still missing from our training data.

Table 5: The Penetration of UTF-8 Encoding in Selected Languages

Language	UTF-8 encoded documents	Document encoded otherwise	Examples of other encodings found [1]
Vietnamese	1,934,392(96.4%)	72,077(3.6%)	TCVN, VIQR, VPS
Mongolian	48,834(95.5%)	2,300 (4.5%)	Latin-Cyrillic
Hindi, Bhojpuri, Magahi, Marathi, Nepali, Sanskrit, Tamang	81,800(78.4%)	22,544 (21.6%)	Agra, Arjun, Kiran, Kruti, Hungama, Naidunia, Shivaji, Shree, Shusha
Sinhala	4,793(44.5%)	5,977(55.5%)	Metta, Kaputa
Arabic	400,933(24.0%)	1,270,189 (76.0%)	Latin-Arabic
Telugu	178(16.6%)	894(83.4%)	Shree, TLH
Tamil	566(14.9%)	3,232 (85.1%)	Amudham, Kumudam, Shree, Vikatan
Hebrew	1,468,344(12.3%)	10,488,970 (87.7%)	Latin-Hebrew
Thai	207,901(2.7%)	7,544,884 (97.3%)	TIS 620
Burmese	24(0.7%)	3,261(99.3%)	WinResearcher
Turkish	20,591(0.5%)	3,938,737 (99.5%)	Latin-Turkish

[1] Local proprietary encodings are shown in this table by names of font (families).

DISCUSSION

In this section, we will discuss issues from the viewpoint of how to realize the vision of an integrated observation-collection instrument for Asian language resources from the Web. First we will discuss the availability and quality of language resources, and then we will focus on our agenda in the technical domains, how to deal with plural scripts and encodings and how to create efficient and workable solutions for collecting language resources.

Overall Assessment

When measured by the number of pages or by pages-per-capita, most of the Asian languages are far less represented on the Web than European languages. This is not a surprising result, but their presence is even more limited than expected. Hindi and Bengali, for example, with almost four hundred million speakers between them – larger than the total population of the European Union – have only one hundred thousand or so pages on the Web. The degree of difference between them and European language representation is in the order of tens of thousands or hundreds of thousands. "The digital language divide" does definitely exist at a worrisome level.

When measured by volume of text, a one million document set contains roughly 5 gigabytes of text, assuming 5000 bytes as the average page size. But only ten Asian languages have above this amount of language resources with the remainder being far smaller.

When we evaluate the quality of documents as language resources, such factors as the variety of content category, language quality, and variety of style of documents should be evaluated. At this moment, we cannot tell much about these points. But at least one point can be mentioned here. It is likely that content category, quality and style of languages are biased, at least in "smaller" languages. The bias might stem from the specialization of usage in a multilingual environment. Multilingualism is the norm in most parts of Asia. In a multilingual environment, there is often specialization in discourse situations. For example, English for the occupational domain is an official language for public or educational domains and other local languages for personal domains. When such specialization is apparent, the language contents on the Web also may show specialization depending upon the domains of the language's specialization. The outstanding dominance of cross-border languages in many country domains suggests that the specialization domains left for the local languages might be relatively limited.

How to cope with the Growing Web

The current survey does not cover Web pages placed under generic domains like com, org or net. Many local language news sites, blog pages and chat-rooms are hosted in generic domains, whose size is almost ten times larger than the entire country code domains. Therefore, considering the growing speed of the Web, the question of how to implement an efficient crawler becomes a key issue in our vision. The current study consumes almost 652 gigabytes of disk storage and consumes 50 to 80 Mbps bandwidth for almost one week. A simple calculation tells us that 65 terabytes of disc storage and 100 weeks would be needed to collect the entire Web (10 billion pages is assumed here). It seems impossible for most non-commercial entities.

In this context, several studies and attempts have been made in the field of language-focused crawling [14][16]. One of the assumptions behind the design of this approach is that pages written in a specific language may have a high likelihood of being linked to pages in the same language. We need to verify this assumption. A graph analysis to reveal the structure of sub-graphs of Web pages written in the same language should be tackled.

In the same context, a distributed crawling approach coupled with proximity-based allocation of tasks has been explored [17]. An advantage of this approach is the possibility of combining free-resources from any possible participant, and proximity-based allocation of tasks can improve the speed of crawling by reducing response time from an assigned server to a target server. The Language Observatory is offering a server to an experiment to test this approach, designed by Thai Computational Linguistics Laboratory (TCL).

CONCLUSION

A detection technique for natural languages and their encoding schemes can also be used as an online language, script, and encoding scheme identifier and to develop tools such as multilingual search engines. It will be difficult to install the shift-codon trained data into a Web browser due to the large amounts of shift-codon required. However, online detection service and crawling for specific language groups could be implemented with limited knowledge, since the server manages the knowledge.

The survey presented, in spite of its limitations, is probably the first comprehensive survey of Asian languages on the Web. The results revealed the existence of a worrisome level of digital language divide and the dominance of cross-border languages in the Asian domains. Through the survey, an estimate of the size of language resources on the Web is given. Also the extent of plurality in scripts and encodings of Asian language documents is indicated. It may be premature to confirm the feasibility of a "Web as Corpus" scenario for Asian languages in a conclusive manner.

Finally, the survey has identified points to be aware of and has given directions that can benefit anybody who tries to create a language resource collection.

Acknowledgement

The study was made possible by the sponsorship of the Japan Science and Technology Agency through its RISTEX program and by the sponsorship of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) through the Asian Language Resource Network project. Moreover, the authors would like to thank Chea Sok Huor (PAN Localization, Cambodia), Valaxay Dalaloy (STEA, Lao), Huynh Quyet Thang (Hanoi University of Technology) and Eedenebat Chuluun (Mongolian University of Science and Technology) for their valuable advice and for providing us with training texts.

References

1. Global Reach, (2006). *Global Internet Statistics*, August 20, 2006, http://global-reach.biz/globstats/index.php3

2. Alis, (1997, June). *Technologies and the Internet Society's survey Web Languages Hit Parade*. http://alis.isoc.org/palmares.en.html

3. FUNREDES report. (2006). *Observatory on the linguistic and cultural diversity of the Internet*, http://funredes.org/LC/english/medidas/sintesis.htm

4. O'Neill E.T., Lavoie B.F., Bennett R. (2003, April). Trends in the Evolution of the Public Web 1998 - 2002, *D-Lib Magazine,* Volume 9

5. Paolillo J., Pimienta D., Prado D. (2005). Measuring Linguistic Diversity on the Internet, *UNESCO Institute for Statistics, Montreal Canada*

6. Mikami Y., Zavarsky P., Rozan M.Z., Suzuki I., Takahashi M., Maki T., Nizan Ayob I. Boldi P., Santini M., Vigna S. *The Language Observatory Project (LOP)*, *www 2005, Proceedings, Chiba, Japan*

7. Caminero R.C., Zavarsky P., Mikami Y. (2006), *Status of the African Web*. WWW 2006, Proceedings, 869-870

8. Boldi P., Codenotti B., Santini M., & Vigna S. (2002). UbiCrawler: A Scalable Fully Distributed eb Crawler. *Technical Report, University degli Studi di Milano, Departmento di Scienze dell'Informazione*

9. Boldi, P., Codenotti, B., Santini, M., & Vigna, S. (2004). UbiCrawler: *A Scalable Fully Distributed*

Web Crawler, Software: Practice & Experience. Vol. 34, No. 8, pp. 711-726

10. Suzuki I., Mikami Y., Ohsato A. (2002). A Language and Character Set Determination Method Based on N-gram Statistics. *ACM Transactions on Asian Language Information Processing,* Vol. 1. No. 3, pp. 270-279

11. Ethnologue. (2005). *Language of the World*, SIL International 2005, 15th Edition

12. Extra, Guus, and Gorter D, (Eds.). (2001). The Other Languages of Europe: *Demographic, Sociolinguistic and Educational Perspectives Multilingual Matters*

13. Bruce P. (1998, May). Turkey and Iran in Former Soviet Central Asia and Azerbaijan: The Battle for Influence that Never Happened, *Eisenhower Institute's Center for Political and Strategic Studies*, Volume 2

14. Pingali P., Jagarlamudi J., Varma J. W. (2006). Indian language IR from multiple character encodings. *WWW 2006, Proceedings,* pp. 801-809

15. Rohra A. and Ananda P. (2005). Collecting Language Corpora: Indian Languages, *The Second Language Observatory Workshop Proceedings, Tokyo University of Foreign Studies, Tokyo*

16. Somboonviwat K., Tamura T., and Kitsuregawa M. (2005). Simulation study of language specific Webcrawling, P*roceedings of the SWOD'05*

17. Tongchim S., Srichaivattana P.,Kruengkrai C., Sornlertlamvanich V. and Isahara H., (2006). Collaborative Web Crawler over High-speed Research Network. *Proceedings, KICSS 2006, Ayutthaya, Thailand*