

Indonesian Languages Diversity on the Internet

Hammam Riza¹, Moedjiono², Yoshiki Mikami³

1: Agency for the Assessment and Application of Technology (BPPT), Indonesia
hammam@iptek.net.id

2: Ministry of Communication and Information, Indonesia
moedjiono@depkominfo.go.id

3: Nagaoka University of Technology, Niigata, Japan
mikami@kjs.nagaokaut.ac.jp

Abstract. The paper gives an overview and evaluation of language resources of Asian languages, in particular of Indonesian official and local languages that are currently used on the Internet. We have collected over 100 million of Asian web pages downloaded from 43 Asian country domains, and analyzed language properties of them. The presence of a language is measured primarily by number of pages written in each language. Through the survey, it is revealed that the *digital language divide* does exist at serious level in the region, and the state of multilingualism and the dominating presence of cross-border languages, English in particular, are analyzed. From this survey as well, the diversity of Indonesian official and local languages on the Internet is observed.

Keywords: Asian language, Indonesian languages, web statistics, language identification, standards, multilingualism, encoding, digital language divide

1. Introduction

Language diversity can itself be interpreted in a number of different ways. Indonesia has more than 740 local languages and India has 427 local languages in its country. Residents of English countries may have many other language skills, but few other countries can match Indonesia for diversity within one country. The numbers of speakers of neo-Latin languages, including those in the US, may be more than twice the numbers of people of English mother tongue but the US controls much of the machinery behind the World Wide Web (Mikami 2005). The relationship between languages on the Internet and diversity of language within a country indicates that even with a globalize network, **nation states have a role to play in encouraging language diversity in cyberspace**. Language diversity can be viewed as much within a country as within the Internet as a whole.

It is a common assumption that English is the dominant force in the Internet. We, as do most others who see English as dominant, view this is a problem. It is reported that English covers about half of all Web pages and its proportion of them are falling as other nations and linguistic groups expand their presence on the Web. Paolillo (2005) points to US dominance of the force behind the Web, both commercial and regulatory, to the extent that the latter exist.

For Indonesia, telecommunication companies who profit from the demand for communication and technology services have a special responsibility to bear in mind the linguistic diversity of the countries whose markets they serve. Hardware and software companies have a similar influence on the linguistic make up of the Internet, by producing computers with keyboards, displays and operating systems that favor particular languages. The acts of computer companies locked in competition for market dominance have a detrimental effect on the climate of multilingual computing and on-line linguistic diversity. In such circumstances, the ethno-linguistic awareness of telecommunication companies, computer companies and Internet governing authorities will begin to broaden only if a critical mass of under-represented ethno-linguistic groups can command their attention. Hence, the general issue of emergent linguistic bias requires close monitoring on global, regional and local scales.

The measurement of languages on the Internet can be used as a paradigm for many issues of measuring content. To put it bluntly if we cannot measure this seemingly simple dimension of

Web site content what can we measure? In this line of thought, we propose the evaluation of Indonesian official and local-regional languages diversity on the Internet.

Measuring the languages in the overall number of pages on the Web increasingly presents challenges caused by the sheer volume of Web content, but just because a page is on the Web does not mean it is used, or even 'visited'. If we are to truly measure the impact of the Information Society, we need to have statistics on how the Internet is used, and by whom. In this view Web pages are simply the supply side, in all its linguistic homogeneity or diversity, and not necessarily a reflection of use and demand. In an oversupplied market of say English language Web pages offering a variety of services, many poor quality sites may receive few or no visits. It is also a common observation that many Web sites remain without updates or modification for years.

Since the early days of web development, various attempts have been made to reveal the language distribution of the web. An estimate of language distribution in terms of the Internet users' language has been regularly reported by a marketing research group (Global Reach, 1996-2005), and estimates of distribution of the web documents are compiled by various groups, each with a different scope and focus. Most of these surveys have evolved along with the development of multilingual search engines like Inktomi, Yahoo, Google, Alltheweb, etc. The language-specific search capability of the search engines has provided means of survey for researchers. Although these surveys have given us fairly good pictures about European language presence on the web, far less attention has been paid on Asian languages, among them "less computerized languages" such as Indonesian local languages in particular.

This ignorance may arise partly from the fact that the "commercial value" of Asian languages has been low, and partly from the technical difficulties of language identification of Asian languages. With the exceptions of Chinese, Japanese, Korean, Thai, Malay, Turkish, Arabic and Hebrew, nothing is known about the extent of Asian languages presence 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*" (Paolillo et al., 2005) 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 (Mikami, 2005). After a few years of development work, LOP team has trained a language identification engine to cover more than three hundred languages of the world, and has acquired the capability to collect terra-byte size web documents from the Internet. This paper is prepared based on the preliminary survey results of LO project with emphasis on Indonesian languages.

2. Objectives

The objectives of this paper are firstly to give an overview for Asian languages on the web, in particular for Indonesian official and local languages which have been ignored up to now. Through this study, we have tried to spotlight the presence of Asian languages as maximum extent as possible. **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 population ratio to give an indication of the relative intensity of web authorship.** In terms of language coverage, we discovered more than fifty Asian languages.

Secondly the paper tries to describe the state of multilingualism in Asian country domains, with special emphasis in Indonesian country domain. The state of multilingualism can be defined at various levels, from a personal or document level to a society 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.

After a brief description on data collection and analytical methodologies, the Asian language presence is discussed, followed by the state of multilingualism in Indonesia and the presence of cross-border languages.

3. Methodology

3.1 WEB PAGES COLLECTED

LOP use a web crawler that works by downloading millions of 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. The collection is also used for various types of web characterization analysis (Caminero, 2006; Nakahira, 2006).

The latest Asia crawl (excluding China, Japan and Korea) focused on web pages in 43 country domains (country code Top Level domain or ccTLD) in Asia. The crawl was begun from a seed file containing 13,286 URLs (see Table 1). Web pages outside of these ccTLDs were not crawled. The crawl was performed by using a decentralized, parallel crawler called UbiCrawler (Boldi et al., 2002). 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.

Table 1: Number of downloaded pages by ccTLD in comparison to Google and Yahoo

Country	ccTLD	Robots.txt found	Number of downloaded / cached pages				
			Language Observatory (LO)	Google ^[1]	LO / Google	Yahoo ^[1] LO / Yahoo	
UAE	ae	125	934,634	4,440,000	0.21	1,140,000	0.82
Afghanistan	af	19	141,261	117,000	1.21	30,000	4.71
Azerbaijan	az	233	2,251,485	2,310,000	0.97	650,000	3.46
Bangladesh	bd	20	207,150	2,840,000	0.07	53,200	3.89
Bahrain	bh	23	246,031	1,410,000	0.17	284,000	0.87
Brunei	bn	5	94,788	1,240,000	0.08	155,000	0.61
Bhutan	bt	9	44,594	233,000	0.19	62,400	0.71
Cyprus	cy	127	627,056	2,440,000	0.26	962,000	0.65
Indonesia	id	1,690	5,742,097	22,100,000	0.26	4,250,000	1.35
Israel	il	18,309	30,943,029	52,300,000	0.59	26,400,000	1.17
India	in	2,156	4,262,378	33,300,000	0.13	8,220,000	0.52
Iraq	iq	0	0	243	0.00	157	0.00
Iran	ir	6,230	4,022,270	7,760,000	0.52	5,070,000	0.79
Jordan	jo	20	287,341	2,200,000	0.13	545,000	0.53
Kyrgyzstan	kg	288	740,921	2,130,000	0.35	348,000	2.13
Cambodia	kh	2	64,265	358,000	0.18	192,000	0.33
Kuwait	kw	4	59,152	2,510,000	0.02	306,000	0.19
Kazakhstan	kz	1,682	6,441,378	3,940,000	1.63	1,670,000	3.86
Lao	la	47	146,635	1,210,000	0.12	256,000	0.57
Lebanon	lb	56	343,538	2,810,000	0.12	1,350,000	0.25
Sri Lanka	lk	37	136,519	1,620,000	0.08	973,000	0.14
Myanmar	mm	1	16,759	445,000	0.04	84,100	0.20
Mongolia	mn	169	400,141	2,660,000	0.15	273,000	1.47
Maldives	mv	6	37,393	414,000	0.09	127,000	0.29
Malaysia	my	1,401	6,865,800	25,900,000	0.27	219,000	31.35
Nepal	np	32	395,901	1,150,000	0.34	481,000	0.82
Oman	om	148	145,207	474,000	0.31	179,000	0.81
Philippines	ph	442	2,732,525	2,480,000	1.10	6,040,000	0.45
Pakistan	pk	82	734,989	4,530,000	0.16	4,060,000	0.18
Palestine	ps	9	88,203	1,390,000	0.06	297,000	0.30
Qatar	qa	10	52,888	985,000	0.05	190,000	0.28
Saudi Arabia	sa	151	1,053,670	6,170,000	0.17	2,120,000	0.50
Singapore	sg	2,856	5,771,191	21,700,000	0.27	221,000	26.11
Syria	sy	5	51,555	632,000	0.08	59,500	0.87
Thailand	th	4,398	12,556,807	38,000,000	0.33	17,100,000	0.73
Tajikistan	tj	19	233,623	219,000	1.07	25,900	9.02

Turkmenistan	tm	23	80,509	255,000	0.32	37,600	2.14
East Timore	tp	714	13,213	178,000	0.07	51,500	0.26
Turkey	tr	2,770	11,363,633	33,900,000	0.34	29,300,000	0.39
Uzbekistan	uz	680	2,286,734	2,710,000	0.84	427,000	5.36
Vietman	vn	341	4,490,288	14,800,000	0.30	5,300,000	0.85
Yemen	ye	3	34,128	115,000	0.30	120,000	0.28
Total			107,141,679	303,065,243	0.35	118,898,357	0.90

^[1] Numbers of Google and Yahoo's cached pages are as of August 8, 2006.

3.2 LANGUAGE IDENTIFICATION PROCESS

Following the downloading process, the language identification engine LIM (Language Identification Module) is used to simultaneously detect the triplet of *language, script and encoding scheme* (LSE is used below for this triplet) for each document. The identification is based on the n-gram statistics of documents. 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.

Languages selected here are official or nationally recognized languages in Asian countries based on the United Nation UDHR data. Table 2 below is the complete list of the Asian languages targeted in this survey, classified by language family. For Indonesian official and local native languages is highlighted in bold. Additional information for the languages is also listed: the script(s) for the language and the encodings we trained.

Table 2: 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
Buginese/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	Kashimiri/Devanagari/UTF-8	Hebrew/Hebrew/Hebrew
Tetun/Latin/Latin1	Hindi/Devanagari/Shree	
Waray/Latin/Latin1	Hindi/Devanagari/KrutiDev	[Turcic]
	Hindi/Devanagari/Hungama	Abkhaz/Latin/UTF-8
[Austroaiatic]	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
	Sinhala/Sinhala/Metta	Thai/Thai/UTF-8
[Mongolian]	Tajiki/Arabic/UTF-8	Zhuang/Latin/Latin1
Mongolian/Cyrillic/UTF-8	Urdu/Arabic/UTF-8	
Mongolian/Cyrillic/8859-5		

^[1] Local proprietary encodings are shown in this table by names of font 8 families

4. Asian languages presence on the web

4.1 INTRODUCTION TO ASIAN LANGUAGES

We can list several language families in 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 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 an official language, and from the mixture of an indigenous language and one of a language, the 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 Middle East Asian countries. Malay in Austronesian language family has around 250 million speaking population in Indonesia, Malaysia, Brunei, Singapore, southern Philippines and Thailand. Dravidian has about 200 million speakers in India, about 3.6 million in Sri Lanka. 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 larger number of speakers with about 125 million and Korean comes with about 75 million.

When we describe the Asian languages, we cannot avoid mentioning the diversity of scripts they use. Contrasted with the US and Europe, the diversity is outstanding. In Southeast and South Asian countries, many scripts which 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.

4.2 WEB PRESENCE BY COUNTRY

The presence on the web of each Asian country is given in Figure 1, where the coloring 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 Web. This map shows that Israel is the highest (4871 pages per 1000 population) in the rank and Singapore and Cyprus follows respectively. The population data was obtained from the CIA World Factbook (estimates as of July 2006).

4.3 WEB PRESENCE BY LANGUAGE

The language identification engine LIM has been trained for more than 200 languages of the world (345 in terms of LSEs) at the time of this survey. Among them, 80 languages are spoken in Asia and the survey found 60 Asian languages among them. The remaining 20 Asian languages are not found at this survey, but note that this does not mean that there are no pages at all for those languages, as the current level of training of LIM is not sufficient and several languages are not yet trained at the time of the survey. Still missing Asian languages from the UDHR listing are Zhuang, Yi, Hmong (including its various dialects), Shan, Karen, Oriya, Divehi, Dzongkha (Bhutanese), etc.

The data shown in the fourth column of Tables 3 show the total number of web pages identified as written in the languages shown in the leftmost column of the table. The data shown in the third column of Table 3 is the speaker population of that language with statistics taken from the UDHR website. The ranking is based on the number of pages. Table 3 shows that Hebrew, Thai, Tatar, Turkish, Farsi, Vietnamese, Malay, Mongolian, Balochi and Javanese have relatively higher 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. Almost similar ranking is observed in both the number of pages and the pages per population.

It can be observed a high degree of “divide” in terms of usage level of languages can be observed even among Asian languages. The number of Hebrew pages per 1000 speakers is 30 times higher than that of the Malay language (ranked tenth in Table 3), 300 times higher than Kashmiri (ranked 20th), and 3,000 times higher than Cebuano (ranked 50th). The speakers’ population of languages is said to follow *Zipf’s Law* - the n-th ranked language speaker is one of the n-th of the population of the top ranked language.

But if we measure the size of language by number of pages written in 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 far progressive power law curve. The situation evidenced here can be well described as a **Digital Language Divide**.



Figure 1. Presence of web pages by Asian countries ccTLD

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
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
Achehese	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
Buginese	Latin	3,500,000	3,533	0.01
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	(100%) (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 as 20% in this survey;

^[2] Almost one third of the pages were found to be an exact copy of another pages. We excluded duplicate pages from language identification process.

5. Multilingualism in the Web

5.1 MULTILINGUALISM BY COUNTRY DOMAIN

The most recent version of Ethnologue (SIL, 2005) lists close to seven thousand languages around the world. More than 2600 of them are spoken in the Asian region. This indicates that huge scale linguistic diversity is observed in Asia. Among 2600, 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 a language of their home use. Official and national language(s) in selected Asian countries is summarized in Table 4.

Table 4: Selected countries with its richest language diversity in Asian region

Country	Number of Languages ^[1]	Country Population ^[2]	Official or National Languages
Indonesia	742	245,452,739	Indonesian
India	427	1,095,351,995	Assamese, Bengali, English, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Marwari, Nepali, Oriya, Panjabi, Sanskrit, Sindhi, Tamil, Telugu, Urdu
China	241	1,313,973,713	Chinese, Zhuang, Uighur, Hmong, Hani
Philippines	180	89,468,677	Filipino, English
Malaysia	147	24,385,858	Malay
Nepal	125	28,287,147	Nepali, Gurung, Tamang
Myanmar	109	47,382,633	Burmese
Vietnam	93	84,402,966	Vietnamese
Laos	82	6,368,481	Lao
Thailand	75	64,631,595	Thai
Iran	74	68,688,433	Arabic, Farsi
Pakistan	69	165,803,560	Urdu, Panjabi, Sindhi, English
Afghanistan	45	31,056,997	Dari, Pashto
Bangladesh	38	147,365,352	Bengali
Bhutan	24	2,279,723	Dzongkha
Iraq	23	26,783,383	Arabic, Kurdi
Cambodia	19	13,881,427	Khmer
Brunei	17	379,444	Malay, English
Mongolia	12	2,832,224	Halh Mongolian
Sri Lanka	8	20,222,240	Sinhala, Tamil, English

^[1] Ethnologue, Language of the World 15th ed. (2005) ^[2] CIA Fact book as of July 2006

Through the survey, the rich diversity of written pages is found in the country with the richest diversity of languages in the region, in Indonesia. **It is interesting to note that there is significantly larger number of pages in Javanese compare to Indonesia.** It is even more surprising if we also include Malay language. Indonesia and Malay language can be categorized into a single root Indo-Malay language spoken in different dialects This is the major language found in Indonesia, Malaysia, Brunei, Singapore, Southern Thailand and Phillipines. The surprising result shows two things: Javanese is dominating web presence in Indonesia and that most of **Indo-Malay websites and pages are hosted in generic domains (.com, net, org etc.) and not in ccTLDs of those countries.** The lesser Sundanese, Madurese, Achehnese and Bugisnese are found to be of great importance to Indonesia's local language diversity on the Internet.

5.2 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. In principle, all Asian languages listed in Table 3 are considered as local languages. The second category is "cross-border languages", such as English, French, Russian, Arabic etc., which are

used as a language of communication among the peoples of different nations. Arabic can be categorized in two ways. In the South East Asia region, English is recognized as an official language in many countries, but also it is working as an important cross-border language. So we treat English in two ways depending on context of analysis. Figure 2 is prepared to show the relative share of these categories of languages in each country domain.

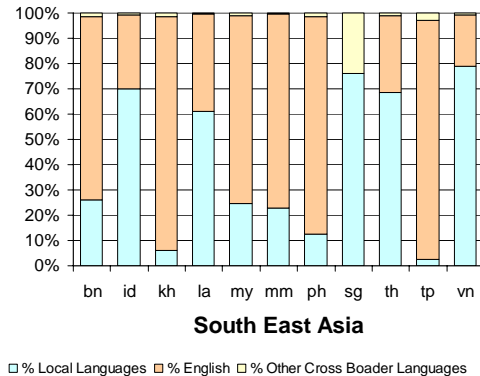


Figure 2: Cross-border languages presence in South East Asia

In South East Asia, the situation is rather different from other sub-regions. **Local languages' share is far higher than in other sub-regions in Asia.** Among them, local language has a majority share in Vietnam (69.8%), Thai (64.0%) and Indonesia (58.7%) in various local languages including Javanese, Achehnese, Sundanese, Balinese, etc. **English dominance is observed and it is reflected in its use on the Internet..**

6. Conclusion

The survey presented, in spite of its limitations, is probably the first comprehensive survey of Asian languages and in particular of Indonesian national and local languages on the web. **The results revealed the existence of a worrisome level of the digital language divide and the dominance of cross-border languages in the Asian domains and in particular, the Indonesian internet domains.**

References

Alis Technologies and the Internet Society's survey Web Languages Hit Parade (1997). Retrieved August 20, 2006, from <http://alis.isoc.org/palmares.en.html>

Boldi, P., Codenotti, B., Santini, M., & Vigna, S. (2002). UbiCrawler: A Scalable Fully Distributed Web Crawler, technical Report, University degli Studi di Milano, Dipartimento di Scienze dell'Informazione.

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

Global Reach, Global Internet Statistics, Retrieved August 20, 2006, <http://global-reach.biz/globstats/index.php3>

Mikami, Y., Zavarsky, P., Rozan, M.Z., Suzuki, I., Boldi, P., Santini, M., & Vigna, S. (2005). The Language Observatory Project (LOP), www2005, Chiba.

Nakahira, K.T., Hoshino, T., Mikami, Y. Geographic locations of web servers. WWW 2006: 989-990.

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

Suzuki, I., Mikami, Y., Ohsato, A. (2003). A Language and Character Set Determination Method Based on N-gram Statistics, ACM Transaction on Asian Language Information Processing, Vol. 1. No. 3, September 2002, pp. 269-278.

Bernard, C., et al.. Atlas des langues, Editions France Loisirs, Paris, 2005

Vikas, O. (2004). Multilingualism for Cultural Diversity and Universal Access in Cyberspace: an Asian Perspective.

Medelyan, Schulz, S., Paetzold, J. Poprat, M, Markó, K. (2006.) Language Specific and Topic Focused Web Crawling. In: Proc. of the Language Resources Conference LREC 2006, Genoa, Italy.

SIL International, Ethnologue 15th Edition,

UNESCO Publication, (2005). “Diversity and Endangerment of Languages in Nepal”, United Nations Educational, Scientific and Cultural Organization, Katmandu Office, Nepal.

UNESCO, (2003). (Adopted by the UNESCO General Conference at its 32nd session Promotion and Use of Multilingualism and Universal Access to Cyberspace”.

Stephen A. Wurm (Eds.) (2001). Atlas of the World’s Languages in Danger of Disappearing. Paris: UNESCO.