

Languages on the Asian and African Domains

Shigeaki KODAMA

Language Observatory Project

Nagaoka University of Technology

1603-1 Kamitomioka-cho Nagaoka-shi Niigata-ken 940-2188 Japan

voice: [+81](258)47-9355; fax: [+81](258)47-9350

e-mail: kodamas@kjs.nagaokaut.ac.jp

Abstract

This article gives results of surveys on language resource of Asian and African languages on the 42 Asian and 48 African ccTLD domains, in particular for those languages that have not been examined so far. These surveys, done by Language Observatory Project, used 100 million pages collected from Asian and 50 million pages from African ccTLD domains in 2006, and one million pages in 2007 each from Asian and from African domains. We found that in both regions, English was the most widely used language, but its presence differed greatly in Asia and in Africa. In Asia the presence, 21.3% in 2006 and 40.6% in 2007, was below the world's average of 45% but in Africa it exceeded the average, with 73.9% in 2006 and 64.9% in 2007. Some of the Asian indigenous languages such as Hebrew, Thai and Turkish were exclusively used in their home countries and showed a considerable presence, but we could not find the same situation among the African languages. In African domains, the presence of European cross-border languages was overwhelming, and the influence of former colonizing countries was observed by the fact that in CPLP (Comunidade dos Países de Língua Portuguesa¹) countries and in Francophone countries, Portuguese and French were the most widely used languages, respectively.

Keywords: language diversity, indigenous languages, Internet, multilingualism, cross-border languages

¹In English, Community of Portuguese Language Countries.

1 Aim of this Article

The aim of this article is to investigate and evaluate the results of LOP research done in 2006 and 2007. The target of the surveys was web pages under ccTLD domains of Asia² and Africa. As these two surveys were done at intervals of one year, we could examine what changes had occurred between 2006 and 2007. In this article, we concentrate on the changes in cross-border languages and the indigenous languages of Asia and Africa. We will examine indigenous languages in detail because, as mentioned below, indigenous languages are not given much attention. In the next section we will summarize the background of our survey, then describe our survey method briefly, and examine survey results.

2 Background of Survey

Language diversity in cyber-space is now one of the most important problems among those who are concerned with multilingualism or multiculturalism in cyber-space. The first step towards determining language diversity in cyber-space is to measure and analyze language distribution. In the UNESCO report presented to the Tunis phase of the World Summit on the Information Society, “*Measuring Language Diversity on the Internet*”[PPP05], we find the same concern.

Since the early days of web, language distribution has been investigated by various organizations and companies. We can find regular reports by Global Reach [Rea06]. Earlier research was done by Alis Technologies and the Internet

²We omitted China, Korea and Japan domains from the targets to avoid mass processing.

Society in 1997[TS97]. FUNDERES compiles a regular report focused on the Romance language group[FUN06]. Those studies targeted mainly European cross-border languages. In those studies non-European languages are largely ignored even if they have many speakers.

It is true that those researches have provided us with fairly good pictures of European cross-border language use, but non-European languages have not been the target of any research, because they are less computerized and less economically valuable. It is clear that when we think of language diversity in cyber-space, research on non-European languages is equally as important as research on European languages. But it is difficult to know the extent of the presence of Asian and African languages, with the exception of Chinese, Japanese, Korean, Thai, Malay, Turkish, Arabic and Hebrew³. There is a strong need to implement an independent survey instrument to observe those languages.

We have developed a Language Identification Machine (LIM) that can identify about 300 languages. We have already reported various survey results at UNESCO's International Mother Language Day and the Internet Governance Forum, and an article on Analysis of Asian languages on the web[ea08] will be published in this year. This article is also a result of our continuous research.

3 Survey Methodology

The procedure of our research is as follows. We first choose seed URLs from which we start crawling. Those seed URLs should be carefully chosen so that our crawler can collect as many pages as possible. Because of the limitation of bandwidth and disk space, our crawler, called UbiCrawler, is configured to stop tracing further links at a depth of 8 and to download a maximum of 50,000 pages per site. While other search engines generally cache all types of files on the Internet, we only collect html and text files.

Collected pages are then processed by LIM, which can identify over 350 triplets of language, script and encoding (LSE) using a statistical

³Global Reach had researched existence of these languages on the Internet in 2004. Now this research is not available but a citation from it can be found at http://en.wikipedia.org/wiki/Global_internet_usage.

method called n-gram. Some LSEs share one or two of their constituents. For example, we have three different LSE entries for the Abkhaz language, in which scripts and encodings are different.⁴ In this article, because we concentrate on language issues, We do not distinguish by LSE but give total values by language We have 75 different indigenous languages of Asia, including Arabic and Chinese, which may be regarded as cross-border languages, and 82 indigenous languages of Africa. Ethnologue[Eth05, p.15] estimates that there are 2269 living languages in Asia and 2092 in Africa. This means that our database covers only 3.3% of them in Asia and 3.9% in Africa, but this fact does not imply that our research is worthless, because the languages stored in our database have a relatively large number of speakers and can cover a majority of the population.

The targets for crawling are 42 Asian ccTLD domains and 48 African ccTLD domains. In 2006, we collected and analyzed about 100 million web pages from Asian domains and 50 million from African domains, and in 2007, we extracted about one million pages each from Asian and from African domains by random sampling and analyzed them. Statistical analysis of the 2006 data proved that a sample size of one million pages represents the entire data with high reliability (over 97%), and therefore this sample size was employed for the 2007 data.

4 Survey results

4.1 Cross-border Languages

We first examine English because it is the most dominant cross-border language in both the real world and cyber-space. FUNREDES estimated in May 2007 that 45% of web pages are in English.⁵ As shown in Table 1, our research in 2007 shows that 40.6% of pages in Asian domains and 64.9% of pages in African domains are in English. As shown in table 1, the presence of English is lower than the average in Asian domains but growth rate is surprisingly high⁶ In African domains, the pres-

⁴They are Cyrillic CP1251, Cyrillic UTF-8 and Latin ASCII.

⁵http://dttil.unilat.org/LI/2007/fr/resultados_fr.htm

⁶The figure shown in Table 1 is the percentage of number of English pages to total number of pages. If we subtract number of pages we can not identify from total, the growth

ence of English is higher than the average, but English is less used in 2007 than in 2006.

Table 1: Presence of English

	2006	2007	growth rate
Asia	21.3%	40.6%	190%
Africa	73.9%	64.9%	87.8%

Among other European cross-border languages, Russian in Asian domains and French in African domains show considerable presence, as shown in Table 2.

Table 2: Presence of Russian and French

	Russian		French	
	2006	2007	2006	2007
Asia	5.1%	3.1%	0.2%	0.3%
Africa	0.1%	0.1%	3.7%	5.5%

More detailed observation reveals that Russian is used mainly in central Asian countries—Kazakhstan, Kyrgyzstan, Uzbekistan, Tajikistan, and Azerbaijan—which were constituent republics of the USSR. Especially in Kazakhstan and Kyrgyzstan, the presence of Russian is remarkably high, at more than 81% in 2006. In contrast, in Turkmenistan and Mongolia, which were under the heavy influence of the USSR, Russian has little presence, less than 2% in 2006. In both of them, English was the most used language on the web and in the latter, Mongolian was used in 29.7% of pages.

In the year 2007, as shown in Table 2, the presence of Russian reduced in Asian domains. Even in Kazakhstan and Kyrgyzstan, Russian became less used and English became more widely used, as shown in Table 3. We cannot simply conclude that English deprive Russian of its share, but we assume that a general tendency for English to be more widely used in cyber-space in Asian domains (see Table 1) can be observed also in Table 3.

In African domains, French was used mainly in countries belonging to the OIF(Organisation internationale de la francophonie) and LAS (League

rate is still 144%.

Table 3: Russian and English presence in Kazakhstan and Kyrgyzstan

	Russian		English	
	2006	2007	2006	2007
Kazakhstan	81.8%	78.7%	7.8%	15.0%
Kyrgyzstan	82.0%	66.2%	8.6%	12.5%

of Arab States) countries. The figure shown in Table 4 reveals that in those countries, French was used far more than the world average of 4.41%⁷.

Table 4: French presence in African domains

	2006	2007
Francophone	35.3%	35.2%
LAS	47.7%	38.1%

We find that in LAS countries, the presence of French decreased by 9.6% between the years 2006 and 2007. The interesting fact is that this reduction is not thought to be a result of expansion of English, but that of Arabic, as shown in Table 5.

Table 5: English and Arabic presence in LAS

	2006	2007
English	31.2%	21.4%
Arabic	9.9%	21.8%

In OIF countries, the presence of English remained unchanged between 2006 and 2007, but that of Arabic increased by 10%. See Table 6 below.

We should notice that six countries belonging to the LAS are also OIF countries. They are Djibouti, Egypt, Comoros, Morocco, Mauritania and Tunisia. When we examine the presence of Arabic in each ccTLD domain of Africa, we find that four ccTLD domains for which the presence of Arabic is greater than 1% belong to both the LAS and OIF, as shown in Table 7. Countries belonging to both the LAS and OIF are marked with bold characters. Notice that in all the countries in Table 7, Arabic is used as an official language and Islam is

⁷See footnote 3 above.

Table 6: English and Arabic presence in Francophone countries

	2006	2007
English	29.7%	28.1%
Arabic	5.0%	15.4%

the most influential religion. Comparing the presence in 2006 with that in 2007, we find that the Arabic presence decreased in Egypt, Chad, and Libya and increased in Tunisia, Morocco, Algeria, and Mauritania. In the ccTLD domains which are not in Table 7, Arabic showed little presence.

Table 7: Arabic presence in African ccTLD domains

	2006	2007	LAS	OIF
Egypt	26.8%	22.2%	yes	yes
Tunisia	16.5%	24.6%	yes	yes
Chad ⁸	4.3%	0.0%	no	yes
Morocco	4.2%	5.0%	yes	yes
Algeria	3.2%	9.5%	yes	no
Libia	2.0%	0.0%	yes	no
Mauritania	0.8%	5.4%	yes	yes

In Asian ccTLD domains, we find that presence of Arabic increased between 2006 and 2007 from 1.6% to 2.6%. Examination of ccTLD domains in Asia reveals that we could divide them into two by presence of Arabic: domains in which Arabic presence is greater than 5% and those in which it is less than 1%⁹. As we have already described for the African ccTLD domains with large Arabic presence, in all of these countries, Arabic is used as an official language and Islam is the most influential religion. These countries also belong to the LAS except for Iran. Arabic presence in these countries is shown in Table 8.

As shown in Table 8, the Arabic presence in the Yemen and Kuwait domains has decreased. And in Palestine, Qatar, Bahrain, the Arabic presence also fell, though the reduction in those countries

⁸In 2007, no collected page is identified as an Arabic page in Chad and Libia.

⁹More precisely speaking, the ccTLD domain coming nearest to Oman is Israel and Arabic presence in the Israeli domain was only 0.15% in 2006.

Table 8: Presence of Arabic in Asian domains

	2006	2007
Yemen	75.5%	35.8%
Kuwait	60.3%	21.6%
Palestine	56.5%	54.2%
Saudi Arabia	50.4%	55.9%
Qatar	48.5%	31.3%
Bahrain	40.3%	32.7%
Syria	37.1%	44.8%
UAE	33.7%	17.3%
Jordan	30.7%	31.5%
Iran	9.9%	12.1%
Lebanon	5.9%	8.4%
Oman	5.5%	6.0%

Table 9: Indigenous Languages in Asian domains

	2006	2007
Asia	52.5%	41.0%
Africa	2.3%	1.2%

was not so much as in Yemen and Kuwait. On the contrary, in the Saudi Arabia, Syria, Jordan, Iran, Lebanon and Oman domains, the presence of Arabic increased in 2007.

4.2 Indigenous Languages

In this section, we will examine indigenous languages in Asian and African domains. We will compare the result of the year 2006 with that of 2007, as we have done in the previous section. And we would also like to try a comparative study of Asia and Africa because this comparison will make it clear what we should do for the promotion of indigenous languages.

The web presence of indigenous languages is given in Table 9. As mentioned above, our identification machine has 75 indigenous languages of Asia and 82 of Africa.

As shown in the table 9, Asian indigenous languages are widely used in Asian domains. On the contrary, the presence of African indigenous languages in African domains was quite limited. If we exclude Afrikaans and Nigerian Pidgin English from African indigenous languages, since they are

of European language origin, the presence become only 0.3% in 2006 and 0.1% in 2007. We notice that the presence of indigenous languages decreased in both Asian and African domains from 2006 to 2007.

In Asian ccTLD domains, seven indigenous languages showed more than 1% presence. These are shown in Table 10 below.

Table 10: Indigenous Languages in Asian domains

	2006	2007
Hebrew	18.5%	17.2%
Thai	12.0%	4.0%
Turkish	6.1%	5.7%
Vietnamese	3.1%	0.4%
Persian	2.0%	2.4%
Javanese	2.0%	1.3%
Indonesian	1.3%	2.1%

We can find an interesting fact when we examine how the pages in these eight languages are distributed: these languages centered in the country domains in which they are official languages. Table 11 below shows the fact. Over 90% of pages in Hebrew, Persian, Thai, Turkish and Vietnamese were concentrated in their home domains.

Table 11: Concentration of Asian Indigenous Language Use in home domain

	2006	2007	Home domain
Hebrew	100%	100%	Israel(.il)
Thai	51.3%	100%	Thailand(.th)
Turkish	98.7%	92.8%	Turkey(.tr)
Vietnamese	99.2%	100%	Vietnam(.vn)
Persian	100%	99.8%	Iran(.ir)
Javanese	81.8%	68.5%	Indonesia(.id)
Indonesian	83.7%	73.6%	Indonesia(.id)

As shown in Table 12, these languages occupied considerable ratios in their home domains.

From these facts, we can conclude that these languages have won their positions in the Internet though they are facing the pressure of cross-border languages like English.

Other major Asian indigenous languages, for

example Hindi or Uzbek, were not so frequently found as the languages discussed above. In the domains in which Hindi and Uzbek are spoken, cross-border languages, English in India and Russian in Uzbekistan, were chosen for over 70% of pages in their domains.

The same situation is found in many of the Asian domains and almost all the African domains. In African domains, we find no African indigenous language whose presence is over 1%.¹⁰ Other major African languages, for example Swahili or Hausa, showed little presence on the Internet though they are used as lingua francas and have large speaker populations.

5 Discussion

As examined in section 4.1, English was the most influential cross-border language in both Asian and African domains. Following this was Russian in Central Asian domains and French in African domains belonging to OIF and LAS. Arabic also was present in the Islam countries. The presence of cross-border languages grew from 2006 to 2007, with the exception of English in Africa.

The dominance of cross-border languages in the Internet, especially that of English, is a reflection of the linguistic situation of the real world and, as Paolillo pointed out in his article[Pao07], the fact that “the Internet technologies themselves are mostly based on English.”

Indigenous languages on the Internet were found to be poorly represented, on the whole.

Table 12: Ratio of Asian Indigenous Languages

	2006	2007	Home domain
Hebrew	62.1%	59.5%	Israel(.il)
Thai	64.2%	55.5%	Thailand(.th)
Turkish	60.7%	65.3%	Turkey(.tr)
Vietnamese	69.8%	48.8%	Vietnam(.vn)
Tatar	–	–	--
Persian	47.8%	47.6%	Iran(.ir)
Javanese	27.7%	19.3%	Indonesia(.id)
Indonesian	19.4%	33.8%	Indonesia(.id)

¹⁰The presence of Afrikaans is over 1%, but as mentioned above, Afrikaans is of a European language origin.

Some of Asian languages are widely used in their own domains, as shown in Table 12, but many of them reduced their presence in 2007 compared to those in 2006. We should conclude that most of indigenous languages which we examined are faced with pressure from these cross border languages. the rapid growth of English in Asian domains and that of French and Arabic in African domains are main factor of the pressure. We also notice that English presence is overwhelming in African domains, though it decreased from 2006 to 2007. there is a difficulty in keeping language diversity against this pressure. The languages we could not examine because of the lack of their samples, is also thought to be under the pressure.

But we may find a clue to promote indigenous languages in cyber-space by examining how and why widely used Asian languages, such as Hebrew or Thai, become utilized in the Internet. A detailed examination will be left for future research, but we would like to point out one probably related fact –code standardization for those languages had been already accomplished before Unicode and they could be handled on computers before the popularization of the Internet.

We could not examine why the growth of cross border languages was caused and the whether the growth is temporal or durable. The latter question may be answered by the next research we are planning to do in 2008, but the former question is a open question and various factors, such as Internet user’s attitude to indigenous languages or the ICT indices like the rate of computer diffusion, should be considered to answer the question.

6 Conclusion

We have examined cross-border languages and indigenous languages of Asia and Africa in Asian and African ccTLD domains. The majority of indigenous languages are under the pressure of cross-border languages, aside from some Asian languages. But many people are now aware of the importance of keeping and promoting language diversity, as UNESCO reports[PPP05], and many organizations are running projects on it. This article gives a short report on the current status of Asian and African languages on the Internet to people interested in the problem. We would like to extend the range of research to South America

and Oceania, and continue periodical surveys in the future to grasp Language choice on the rapidly changing Internet.

Acknowledgments

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.

References

- [ea08] S.T. Nanadasara et al, *Analysis of the asian languages on the web based on n-gram language identification*, The International Journal on Advances in ICT for Emerging Regions (2008).
- [Eth05] *Ethnologue*, SIL International, 2005.
- [FUN06] FUNDERES, *Observatorio de la diversidad lingüística y cultural en la internet*, <http://funredes.org/LC/english/medidas/sintesis.htm> (2006).
- [Pao07] John C. Paolillo, *How much multilingualism? language diversity on the internet*, The Multilingual Internet: Language, Culture, and Communication Online, 2007.
- [PPP05] J. Paollio, D. Pimienta, and Daniel Prado, *Measuring linguistic diversity on the internet*, UNESCO Institute for Statistics (2005).
- [Rea06] Global Reach, *Global internet statistics, august 20,2006*, <http://global-reach.biz/globstats/index.php3> (2006).
- [TS97] Alis Technologies and Internet Society, *Web languages hit parade*, <http://alis.isoc.org/palmares.en.html> (1997).