

Extending Our Sense of Cyberspace Language Plurality: The Value of the Language Observatory (LO) Project

Mohd Zaidi Abd Rozan* and Yoshiki Mikami**
Nagaoka University of Technology
Niigata, JAPAN
940-2188

Abstract

As the World Wide Web (WWW) grows exponentially, multilingual web pages are flooding the cyberspace at a tremendous rate. Most probably many of us would guess that the main medium of language on the Web is English. On the contrary, according to greach.com [1], there are 801.4 million people online with at least 510 million non-English and the remainder are English speakers. As a big step towards comprehending web page dimensions regarding languages in cyberspace, we have officially launched a project called "Language Observatory (LO)" in February 2004. We have made several experimental runs using Ubicrawler, some of which were dedicated to the 57 Organization of the Islamic Conference country code Top Level Domains (ccTLD). It is interesting to note that we covered at least 42 million web pages compared to almost 17 million indexed by two well known search engines and this covers nearly triple the amount containing multiple dimensions such as languages, script and character set encoding. Furthermore, data mining activities by LO yield significant findings that further provide a snapshot of cyberspace. This will offer contents that are often created in particular domains hence this provide practical information: language

preferences and source documentations in cyberspace. The potency of LO in producing indispensable information must be taken into account because these are factors that should not be absent within the value chain of translation activities.

Keywords: Language Observatory, Web pages, Language Scenes, Web Intensity, Translation, Language, Script, Character set, Crawler, Language Digital Divide

1. Introduction

With the proliferation of Internet technology, today's emerging knowledge society uses "information" as its recipe for better decision-making and empowerment. As the ability to access information on the Web is very much dependent on the language used for it, this has created disadvantages for groups that are not able to access information that is not viewable in their language. Furthermore, the current estimation of spoken languages around the globe today is around six to seven thousand. If we visit the site of the Office of the Higher Commissioner for Human Rights of the United Nations, we find more than three hundred different language versions - from Abkhaz to

* Graduate School of Information Science and Control Engineering and also corresponding author. *E-mail address:* zaidi@mis.nagaokaut.ac.jp

** Professor, Management and Information System Science Department

Zulu - of the "Universal Declaration of Human Rights" [2]. But we also find that many language translations, especially non-Latin script based languages, are just posted as "GIF" or "PDF" files, not in encoded texts. This situation can be described as "digital divide among languages" or just "language digital divide". As a UNESCO resolution mentions [3], "the promotion and use of multilingualism and universal access to cyberspace" is an urgent item in the agenda of the global information society.

In this paper we will provide information on several language usages in the Organization of the Islamic Conference (OIC) cyberspace as retrieved from Google and also its lack of specifics for language filtering. This is followed by the rationale of LO project as well as its role, vision and the LO crawl for OIC nations.

2. Language Scene in Organization of the Islamic Conference (OIC) countries



Figure 1. Trilateral Chart with 3 corners of languages

2.1. Multilingualism of OIC Countries

As a glimpse of the web, we decided to use Google to get results for specific languages in a country domain. This has produced exciting findings regarding the utilization of language currently in the cyberspace. In Appendix 1, we list 4 major

languages that recorded high returns and the countries are categorized geographically. Based on this, we developed Figure 1 and we coined it as the "Trilateral Chart". The main function of this chart is to create a visual representation of the various scores shown in Appendix 1. This Trilateral Chart has 3 corners that

represent English (top), Arabic (bottom right) and French/Russian (bottom left). The language that is more in use in each country is shown by the shorter distance to the corners of the triangle.

Contained in this chart are points representing 56 OIC countries[#]. For example, Tunisia is located midway between the Arabic and the French/Russian corner and has a longer distance to English corner. From this, we can say that French/Russian and Arabic are equally dominant but English is minor in Tunisia.

Again, by looking at Figure 1 a few points can be easily observed. Firstly, Gulf countries like Saudi Arabia, Oman, Qatar, Kuwait and Bahrain are found to be the heaviest users of Arabic. This is also noticeable for Syria, Yemen, Sudan and Palestine.

Secondly, most Asian OIC countries are not using Arabic on the web to a great extent; English is mostly used as an international language instead of Arabic and French. This is recorded for countries such as Malaysia, Bangladesh, Indonesia, the Maldives and Pakistan.

Thirdly, Francophone members of OIC countries are still writing mostly in French, but Lebanon, Chad and Djibouti are using English more than French. Togo is also not using French very frequently.

Lastly, the former United Socialist Soviet Russia (USSR) states are still employing the Russian language. This comprises countries like Kazakhstan, Tajikistan, Uzbekistan, Kyrgyzstan, Azerbaijan and Turkmenistan. The domination of the Russian language is still very much influencing the cyberspace of these countries.

[#] Excluding Iraq: no results gained from this search

Although the Trilateral chart at this time has been created only for OIC countries, it is possible to further the comprehensiveness by adding all 247 available country code Top Level Domains (ccTLDs) for all the countries on the globe. An added advantage can be achieved if the relation towards the national language(s) usage is also drawn.

2.2. Inadequate Recognition of Languages

2.2.1. No specific differentiation for local Language

As the default, Google provides return results in 'any language'; however, its Advanced Search provides return results in 35 languages. This advanced feature is very much helpful only if we are not trying to differentiate explicitly a language – for example, Arabic. Filtering Arabic language to the breakdowns of its family groups such as Dhofari, Baharna and even 20 other Arabic versions is not possible. As for Indonesian pages returned by Google, it includes a huge number of pages written in Malay from Malaysia and also Brunei. This shows that there is no distinction made between the Indonesian and Malay languages. Such mixes of language are contributing to the low accuracies of search results generated.

2.2.2. Lack of Language Specific Search Engine

A search engine that is capable of identifying language type is very much lacking in cyberspace. Google and other available search systems are competent at meeting demands from users, as long as the query is done within the list of supported (but sometimes undifferentiated) languages. And overall, Google gives a very good picture of global languages. But when national language is involved, they are

scantly represented. This has been one of the reasons why search engines are not the most reliable source of information when language is the matter under concern.

3. The role and vision of Language Observatory (LO)

The pure motivation for this project was the imbalanced usage of language in cyberspace. Having recognized such an urgent challenge, the Language Observatory project was planned to provide a means for assessing the usage level of each language in cyberspace. More specifically, the Language Observatory [4] is expected to periodically produce a statistical profile of language/scripts/character code (LSC) usage in cyberspace. Once the observatory fully functions, the following questions are to be answered: How many different languages are found in the virtual universe? Which languages are missing from the virtual universe? How many web pages are written in any given language, say Pashto? How many web pages are written using the Tamil script? What kinds of character encoding schemes (CESSs) are employed to encode a given language, say Berber? How quickly is Unicode replacing conventional locally developed encoding schemes on the net? Along with such a survey, the project is expected to work on developing a proposal to overcome this situation both at a technical level and at a policy level.

4. Rationale for LO

As mentioned in the previous section, many questions are to be answered in the process of realizing our vision. Our activities cover a wide range of tasks within the progression of the project. At this time we are developing a Language Identification

(LI) module based on N-gram algorithm [5] together with the development of N-gram teacher data from many languages. Some other activities are language database development, character database development, language-relation analysis through linkage structure (dialect family or mutually communicable languages between pages) and many more. We believe that none of these needs will be covered effectively by any web search engines, and therefore we need to develop our own tools to extend the senses for language plurality in cyberspace.

5. LO: First Crawling Experience

5.1. Initial Crawling Target

A group of 57 Islamic countries, members of the Organization of the Islamic Conference (OIC), were chosen for the Observatory's first-round experimental data collection using Ubicrawler [6]. Needless to say, the selection does not imply any political or religious interest of the project. OIC countries not only cover Arabic speaking countries but also cover non-Arabic speaking countries like Malaysia, Afghanistan and Pakistan. The latter group of countries uses Arabic script together with their own extensions, which sometimes cause trouble with text encoding. OIC countries also cover several non-Arabic script users such as Turkey, Tajikistan, and Kyrgyzstan.

5.2. Coverage Performance

Our method has successfully produced a relatively huge amount of data from the web. The number of 57 OIC ccTLDs recorded greater results than with Google, with 42.9 million files retrieved compared to only 5.9 million returned by Google. The distribution of coverage in terms of

downloaded HTML pages count reached over 7 times more than Google's. We also did a similar comparison to Yahoo (10.9 Million returned), and our page count is nearly 4 times higher in the form of

searchable HTML pages (*Note: Returns by Google and Yahoo search engines were on 07/02/2005*). Comparison of Google and the LO in terms of coverage is shown in Table 1.

Table 1. Comparison of Google, LO and Seed URLs in selected countries

Countries	ccTLD	HTML Documents (x 1000)			
		*Google Hit Page	(A) LO Downloaded Page	(B) Seed URL Prepared	Ratio= (A):(B)
Turkey	.tr	2,200	9,923	100	99230
Malaysia	.my	1,170	9,332	101	92396
Indonesia	.id	567	6,538	95	68821
Kazakhstan	.kz	284	4,126	100	41260
Uzbekistan	.uz	97	2,252	100	22520
Pakistan	.pk	237	1,355	100	12550
Saudi Arabia	.sa	109	1,338	100	13380
United Arab Emirates	.ae	63.6	913	98	9316
Iran	.ir	81.7	913	100	9130
Azerbaijan	.az	134	874	100	8740

**Accessed on Feb 07, 2005*

5.3. Uniform Resource Locator (URL) Seeds

The preparation of URL seeds for the crawl was not a complicated task but rather a laborious process. For this pilot test, all the seeds were processed from returned pages of Google. The URLs were not unique though; a manual process of eliminating duplication was done before listing an average of 100 URL seeds for each ccTLD. Some returned results were very poor, such as ccTLDs for Comoros (.km) and Guinea-Bissau (.gw): only 22 and 7 respectively. The seed URLs were then submitted to the crawler for action. One seed URL actually represents 1 page or more probably multiple framed pages that contain hyperlinks. These hyperlinks are then trailed by the crawler agents exhaustively until there is no linkable page found. While in the process of trailing, all possible html pages will be fetched and stored in a chunk form in a repository in our server machines.

5.4. Web Intensity/ Connectedness

The scenario of intensity or connectivity in the web can be understood. From our observation it was found that the selection of URL seed pages is a determinant of the coverage by the crawler agents. If the web is connected well, even a single point is usually adequate to make sure that the crawler exhaustively crawls to every site. However, if it is not connected well, we need many entry points for it to obtain better coverage. In the real world, virtually not all web graphs are connected, this means that they have dead-ends, islets, etc. Our idea is that at least a few seed URLs should cover different islets in order to establish crawling path at minimum to their own islets, and this would probably produce better quantity of downloadable pages by the crawler. Table 1 also shows the relationship between seed URL prepared and the downloaded pages by LO in terms of

ratio. The bigger the ratio, the better linked they are to any targeted pages. In this case, the more connected the web is, the larger the language asset that can be found. Information such as subject area, industry sector and language opportunity would be the list of the supporting value of translation initiatives.

6. Conclusion

In the cyberspace where most information is scattered in a distributed way, a reliable method of obtaining specific information is badly needed. With the limited capability of search engines to distinguish language properties of documents, it is surely important to establish another method in order to cover national or local languages. After having done several experimental runs in LO, we believe that our approach is likely to produce better coverage and further improve our data for later mining purposes. Such definite way of retrieving language related information will provide a better understanding of the setting and also contribute to the higher productivity of messages conveyed through effective translation. It must be kept in mind that activities like marketing, branding, and customer service rely greatly on the precious language assets that are produced and this could generate more perfect communication with added value.

Acknowledgement

The Language Observatory Project is funded by Japanese government through the Japan Science and

Technology Agency (JST) under the RISTEX Program and is supported by UNESCO and Japanese National Commission for UNESCO. The authors would like to thank the Japan Science Technology Agency (JST) for the financial support provided in making this paper presentation a reality.

References

- [1] <http://www.glgreach.com/globstats/> Accessed on 28/12/2004
- [2] <http://www.unhchr.ch/udhr/navigate/alpha.htm>. Accessed on 8/2/2005
- [3] http://portal.unesco.org/ci/en/ev.php-URL_ID=13475&URL_DO=DO_TOPIC&URL_SECTION=201.html Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace, 2003. Accessed on 8/2/2005
- [4] Y. Mikami, P. Zavorsky, A. R. M. Zaidi et. Al, *The Language Observatory Project (LOP)*. Accepted Poster Paper for The 14th International World Wide Web Conference (WWW2005), Chiba, Japan.
- [5] I. Suzuki, Y. Mikami, A. Ohsato, Y. Chubachi, *A Language and Character Set Determination Method Based on N-gram Statistics*, ACM Transactions on Asian Language Information Processing, 1(3): 270-279, 2002.
- [6] P. Boldi, B. Codenotti, M. Santini and S. Vigna. *UbiCrawler: A Scalable Fully Distributed Web Crawler*. Software: Practice & Experience, 34(8):711-726, 2004.

Appendix 1 : Language usage in the OIC Cyberspace

No	Countries (Geographical Location)	*Google total number of results			
		Arabic	English	French	Russian
	MIDDLE EAST				
1	Afghanistan	65	7,280	5	0
2	Bahrain	56,300	34,500	27	2
3	Iran	1,350,000	438,000	1,290	13,000
4	Iraq	0	0	0	0
5	Jordan	150,000	92,000	41	3
6	Kuwait	69,800	25,400	12	1
7	Lebanon	17,400	372,000	41,000	544
8	Oman	51,300	15,400	6	0
9	Pakistan	73	1,110,000	96	486
10	Palestine	132,000	35,800	43	124
11	Qatar	33,400	11,500	1,130	0
12	Saudi Arabia	695,000	172,000	256	11
13	Syria	5,670	362	4	0
14	Turkey	1,980	767,000	13,400	11,800
15	United Arab Emirates	369,000	405,000	3,410	161
16	Yemen	21,700	3,680	9	0
	CENTRAL ASIA + CAUCASIAN				
17	Azerbaijan	0	90,700	64	248,000
18	Kazakhstan	2	165,000	221	1,530,000
19	Kyrgyzstan	33	48,300	224	249,000
20	Tajikistan	0	479	1	4,440
21	Turkmenistan	4	11,400	81	23,100
22	Uzbekistan	8	129,000	39	712,000
	SOUTH ASIA				
23	Bangladesh	0	108,000	7	0
24	Maldives	14	16,300	2	5
	SOUTHEAST ASIA				
25	Brunei	18,100	34,400	4	62
26	Indonesia	58	567,000	406	51
27	Malaysia	1,270	1,440,000	1,140	249
	AFRICA				
28	Algeria	30,300	8530	84,500	1
29	Benin	0	373	2,260	0
30	Burkina Faso	0	854	32,900	1
31	Cameroon	0	1,350	9,370	0
32	Chad	13	274	99	0
33	Comoros	0	7	9	0
34	Cote d'Ivoire	0	383	27,700	0
35	Djibouti	6	13,700	9,350	1,350
36	Egypt	348,000	419,000	6,030	4
37	Gabon	0	119	573	0
38	Gambia	0	4,160	8	0
39	Guinea	0	96	907	0
40	Guinea-Bissau	0	12	0	0
41	Libya	135,000	278,000	1,500	145
42	Mali	0	78	7,490	0
43	Mauritania	5,210	894	12,400	0
44	Morocco	26,100	30,900	499,000	12
45	Mozambique	0	8,470	86	0

46	Niger	1	2,400	2,720	4
47	Nigeria	0	7,070	1	0
48	Senegal	4	5,430	95,600	1
49	Sierra Leone	0	213	0	0
50	Somalia	0	12	0	0
51	Sudan	1,390	331	0	0
52	Togo	11,700	3,660	1,500	0
53	Tunisia	163,000	34,300	166,000	4
54	Uganda	1	71,000	269	16
	EUROPE				
55	Albania	1	48,000	9	2
	SOUTH AMERICA				
56	Guyana	0	3,120	0	0
57	Suriname	77	2,430	0	2,080
	Total	3,693,980	7,045,667	1,023,199	2,796,659

**Accessed on March 27, 2005*