Web-Based Vulnerable Peoples
– Focusing on Language –

Katsuko T. Nakahira
Nagaoka University of Technology

1 Introduction

In these day, human communication is very active in the Internet. Various types of human communication take place on the Internet, have been evolving day by day, including uni-directional distribution of information by means of Web contents such as blogs and bi-directional exchange of information via social media.

Since "language" is the primary media for representing information on the Internet, and "language" is manipulated both by people and the computer systems that control information flow on the Internet, some types of "divide" should result due to the discrepancy between the languages respective peoples want to speak and read for distributing and acquiring information, and those the computer systems are using for transmitting and receiving digital signals that convey the information.

This communication focuses on this discrepancy and the digital divide it should cause, which ultimately yield "linguistic originated vulnerable peoples." This paper addresses the linguistic originated vulnerable peoples issue by introducing a human-centered perspective, which provides a basis for considering the mechanism under which the linguistic originated vulnerable peoples are produced and observing how the linguistic originated vulnerable peoples exist on the Internet.

There are several types of communications; for example, asynchronous communication via blog or web pages, bi-directional human communication with social media. These communications are regarded as the result of peoples' activities of converting ones' thoughts, which originate mainly from their primary languages, into information on the Web expressed by the coded characters.

We discuss the ecology of linguistic originated vulnerable peoples [1] and present new dataset of LOP(Language Observatory Project,Mikami et al. (2005)).

2 Two Types of Language: Language 1 and Language 2

An *ordinary language* that a people uses daily can be classified into two types; the first type of language, Language 1, is the one that the people wants to use, and the second type of language, Language 2, the one that the people is forced to use. Language 1 is associated with the people's identity and used with the people's intention.

On the other hand, Language 2, otherwise called "official language", is associated with politics that govern over the peoples. It is often used among multiple peoples having their respective own Language 1 for the purpose of establishing mutual communication. However, since Language 2 should not be an ideal communication medium for those peoples who use different Language 1 than the Language 2 as their ordinary language, they tend to have a limited ability to express what they want to say in the Language 2.

Non-professions, or ordinary computer users, have increasing opportunities to express their ideas produced by their creative activities via their own language, Language 1, on the Web. The Web may or may not be Language 1 friendly. In other words, the computer systems that reside underneath the Web would either:

1. take care of Language 1 without any problem, or
2. manage to handle it with some inconvenience on the part of the users, or
3. will not accept it at all.

By focusing on Language 1, we can gain important insights into how various types of digital divide might occur on the Web; the Web however must be perceived as an ideal arena for those who want to make full use of it via *their* language. We call this approach "human-centered perspective on digital divide." In the following, we show the main elements of human-centered perspective on digital divide. It turns out that they are very effective to identify

[1] Note that the word "people" is used in this paper as a singular noun with the meaning *the people who belong to a particular country, race, or area*

the kinds of *potential* digital divide that should occur irrespective of the size of the population of Language 1 people.

3 Definition of Linguistic Originated Vulnerable Peoples

Continuing on the above-mentioned argument, we define "linguistic originated vulnerable peoples." Suppose that there is a *language* which has script, like Arabic script, or it is represented by transcription, in other words, it is associated with a symbol system, and it is spoken or written by a specific people characterized by the country it belongs to, the race, or the area where it lives. This defines the *ordinary language* that a specific people uses.

Then we can define two types of vulnerable peoples on the Web as follows: The first is the language speakers who cannot represent their Language 1 in the symbols on the computer systems underlying the Web, or language readers who cannot convert the symbols on the computer systems to their Language 1.

The second is the language speakers who can represent their Language 1 as symbols on computers without any problem, can make easy access to the Internet, but cannot have a free press due to political, social, and the other reasons. It means that the Language 1 speakers cannot create contents faithfully representing their intention.

4 Novel Approach for Dealing with Linguistic Digital Divide

For the Web users, such a situation as mentioned above has influence on the flexibility in creating Web contents. On the Web, any person who wants to send her/his messages as the Web contents is supposed to be a creator regardless of her/his computer skill. It means that the richness of the contents in a certain language can crucially depend on whether the creators of contents can easily manage that language, i.e., Language 1, in computer systems or not.

This issue is directly related to linguistic diversity on cyberspace($\sim$ Web). However, when we discuss the linguistic originated vulnerable peoples in connection to the digital divide and the Web contents, its difficulty and sociality prevented us from treating the relation as engineering. The Web contents, which are objective observables on the Web, may be influenced by a certain degree of digital divide, if there is any, and therefore, revealing what is happening in

back of the Web contents have been relying on inductive reasoning techniques or questionnaire surveys worked out by social organizations. We believe these approaches have potential limitations; inductive reasonings could generate some hypothesis to be proved and it is difficult to distribute questionnaires to the wide range of peoples on the Web. We think that this limitation of the approach to this issue is one reason why the discussion does not reach the depth that clarifies the mechanism of yielding linguistic originated vulnerable peoples . To understand this mechanism, we introduced the framework of e-Network, which presented last NovemberNakahira (2012). The framework suggested that the most important component for describing *digital divide* is media – in this case, ***language***: connecting all components of framework, human factor, substratum factor, and product.

Based on the framework, we can easy understand the relation of linguistic originated vulnerable peoples and digital dividel. To indicate the phenomenon, the first step is to collect languages used in Web. The purpose of LOP is very fit for the study, and we already have present the language diversity in cyberspace at 2006, Asia(Nandasara et al. (2008)). In this communication, we would like to introduce the ***preliminary*** data at 2012 Asia.

5 The survey

Our survey is based on Nandasara et al. (2008), excepted based seed URL. We treated 51 country domains in Asia, including CJK. The crawl was begun from a seed file containing 825 URLs which selected from portal site. The list of ccTLDs contains ae, af, am, az, bd, bh, bn, bt, cn, cy, eg, ge, hk, id, il, in, iq, ir, jo, jp, kg, kh, kp, kr, kw, kz, la, lb, lk, mm, mn, mo, my, mv, np, om, ph, pk, qa, sa, sg, sy, th, tj, tl, tm, tr, tw, uz, vn, ye. The crawl was operated in mid Sep. 2011 to early Jan. 2012. We downloaded about 3.2×10^7 pages without duplication.

There are two difference between two crawling conditions. The first difference is the colelcted seed URLs. 2006 crawling seed URLs are collected special site. 2012 crawling seed URLs are mainly collected portal site, 5 seed URLs per ccTLD so that there are less seed than 2006. The second difference is whether CJK ccTLD is included or not. 2006 crawling is NOT included, but 2012 crawling IS included.

Figure 1 to Figure 3 show some preliminary results. Figure 1 shows percentage of web pages on the Internet as ccTLD level, compared with 2006 and 2012

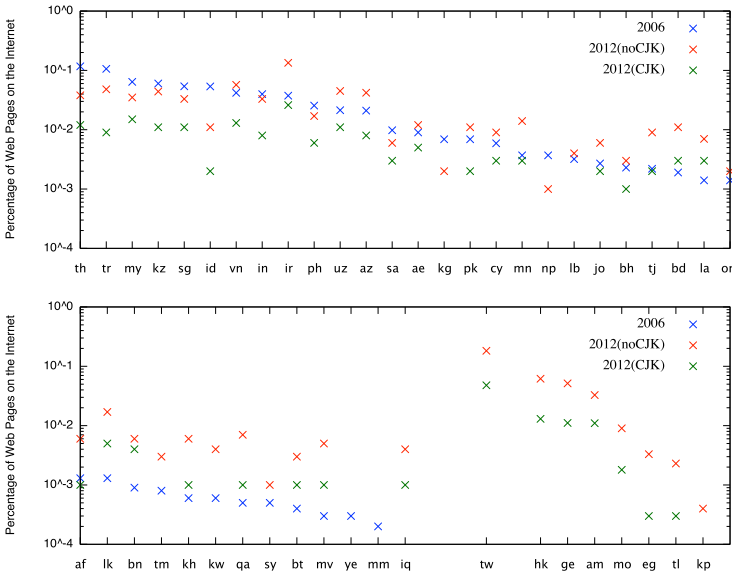

Figure 1: The Distribution of Percentail of Webpages, Compared with 2006 and 2012 Asian Crawling.

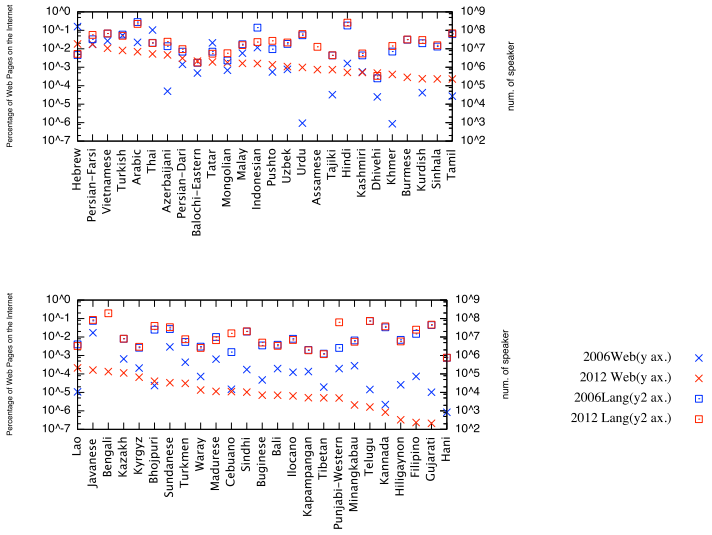

Figure 2: The Distribution of Percentail of Language in Webpages, Compared with 2006 and 2012 Asian Crawling.

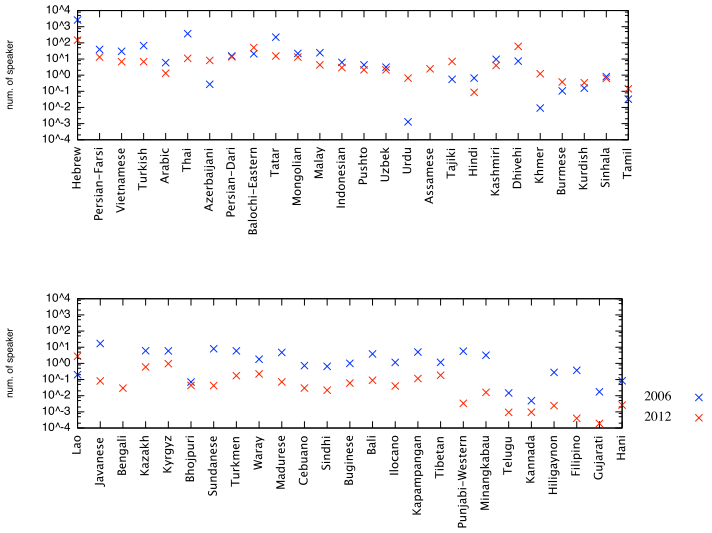

Figure 3: The Number of Webpages per Speaker, Compared with 2006 and 2012 Asian Crawling.

Asian crawling results. 2012 crawling trends are almost same as 2006 crawling, but there are several shifts. The percentage of **id, th** collected webpages are less than 2006, but almost percentages of collected webpages per ccTLD are growth. Especially, **ir, mn, tj, bd, la, af, lk, bn, tm, kh, kw, qa, bt, mv** are growth up to 4 times than 2006 crawling.

Figure 2 shows percentage of language pages on the Internet at Language level compared with 2006 and 2012 Asian crawling results, including the Language speakers number for each years. Left vertical axis is percentage of web pages on the internet per the language, which represented by $\times$. Right vertical axis is the number of the speaker per the language, which represented by square. The blue color is the result for 2006 crawling, and red one is as for 2012 crawling. The dataset sorted by 2012 web pages on the internet per the language. From the figure, percentage of web pages on the internet per the language which represented in underside the graph seems increasing. We guess there are several reasons for increasing – the difference of the aim of crawling, time effect, and so on –. But anyway we need to make several focused crawling if we clarify the phenomena.

Figure 3 shows percentage of web pages per language speaker. The 2012

crawling trends are almost same as 2006 crawling results, excepted the Urdu increasing and Punjabi-Westen decreasing.

The author thinks the trends is caused by focused crawling, namely collecting seed URLs by portal sites, but we need more research. In case of special seed URLs, the page generator might think they feel several mind of archiving or preserving for the language. But in case of portal site seed URLs, the contents generator only behave natural so that the language they use is their natural language, namely Language 1. The growing the cyberspace world, the Internet user can easy access and generate contents. If we would like to know human's natural behavior for language, we might to make crawlinng to portal site.

6 Future Work

In this communication, we show a digest and preliminary report of Asian crawling done at 2006 and 2012. And now, we just done Carribean 2012 crawling, which will be analysed by Daniel Pimienta. For future, we would like to make a plan for other minolity language observation, especially Russian. If you have any information for Language dataset, such as Language name, ISO 639 code, Availability of UDHR text, and useful seed URLs, please do not hesitate to contact to us.

References

Mikami, Y., Zavarsky, P., Zaidi, M., Rozan, A., Suzuki, I., Takahashi, M., Maki, T., and Ayob, I. N. (2005). The language observatory project (lop). In *In Special interest tracks and posters Proceedings of the 14th international conference on World Wide Web*, pages 10–14. ACM Press.

Nakahira, K. T. (2012). Framework for understanding human e-network — interactions among language, governance, and more. In *3rd International Symposium on Multilingualism in Cyberspace(Paris)*.

Nandasara, S. T., Kodama, S., Choong, C. Y., Caminero, R., Tarcan, A., Riza, H., Nagano, R. L., and Mikami, Y. (2008). An analysis of asian language web pages. *The International Journal on Advances in ICT for Emerging Regions*, 1(1):12–23.