

The Language Observatory Project and its Experiment: Cyber Census Survey

Yoshiki Mikami and Izumi Suzuki, Nagaoka University of Technology, Japan

What we are going to do?

The Language Observatory

A web information center (the Language Observatory) that provides online text data to researchers will be founded. More than a hundred of web crawler PCs and a large-scale data storage are going to be equipped here, likewise a commercial search engine. It will also allow researchers to use their own crawler robot on the observatory.

The Cyber Census Project

The highest priority experiment of the Language Observatory is the Cyber Census project, which surveys the activity levels of each language. More specifically, it observes how many web pages exist on the web by language, script, and character set, every year for at least ten years.

What the purpose? --- The Background

The Language Observatory

It looks as though a commercial search engine can be useful for getting such the text data, however, 1) they refuse such large access, 2) they do not distinguish the difference between upper and lower case, and 3) they do not search for ASCII symbols such as “#” and “&”. That is why a language observatory that can trace most of the web effectively is now requested by researchers who survey various activities around languages on the web.

The Cyber Census Project

(1) Raise awareness on the Digital Divide between Languages

There are many kinds of statistics showing the digital divide between specific people, though let's look at the global digital divide between languages and writing systems. A statistic [1] shows that the online population is estimated to be 390 million, in total. Of all these people online, 48% (216 million) use English, 29% (131 million) use non-English western European languages, 9% (40 million) use Chinese, and another 9% use Japanese. No more than 10 million people use a language other than those listed above in their online community.

An even more serious issue is the digital divide between writing systems. **Figure 1** shows the distribution of global population, writing/printing paper consumption, and online population. Latin alphabet users, 39% of the world's population, consume 72% of the writing/printing paper, and occupy 84% of the online population. On the other hand,

Indic, Arabic, and Cyrillic script users consume considerably less writing/printing paper and Internet resources compared with their population.

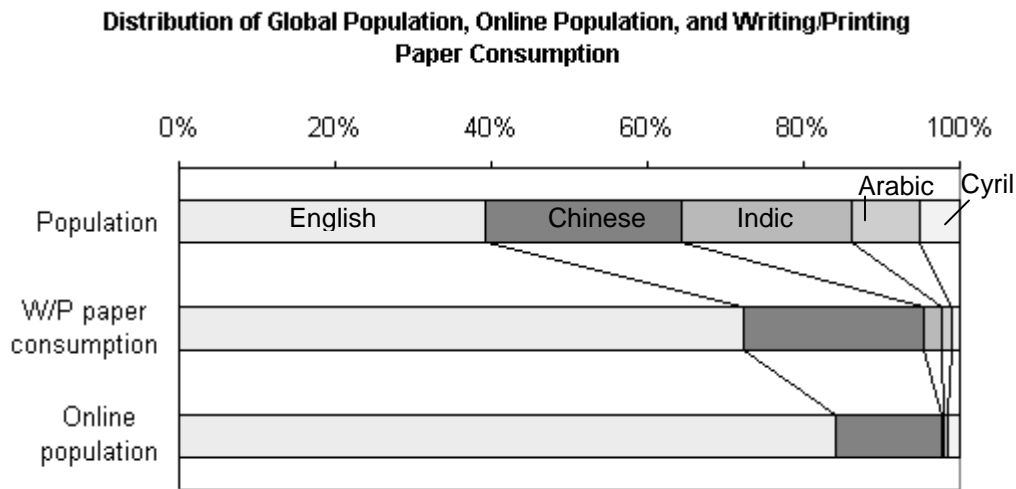


Figure 1 Source: Mikami [2]. Chinese includes Japanese and Korean. Indic includes all the Brahmi-origin scripts such as Myanmar, Thai, Lao, Khmer, and every language in India.

(2) Watch the rate of Unicode-employed pages for each language

Unicode is not effectively used in many languages in and around India. Instead each font designer adopts his or her own encoding scheme, which creates an obstacle for local people sending e-mail and browsing websites. This is the one of the factors that causes the digital divide between languages. The real situation will be unveiled as a result of this project, and also the rate of Unicode-employed pages can be observed every year.

(3) Provide Indian languages online corpus useful to regularize the set of combined character to be used in Unicode texts

In addition to that, an online corpus of languages in and around India will be provided through the project. In these languages, vowel and consonant signs are graphically combined to form characters. The shapes of the original signs usually change when they are combined. As long as Unicode provides only the vowel and consonant signs, we must decide the encoding rules governing the encoding order of vowel and consonant characters. However, in some languages, neither the encoding order rules nor the set of entire characters to be dealt with in Unicode has been declared yet. This project can provide a well-balanced online corpus in these languages.

(4) Which language is missing in the virtual universe?

Existing Survey

Global Reach <http://www.greach.com> provides “Global Internet Statistics” on its website, which is an estimate of the number of people online in each language zone. The NUA (ComputerScope Ltd.) <http://www.nua.com> has surveyed how many people are

online throughout the world. However, no information has as yet been reported about the actual or estimated number of web pages by language, script, and character set.

Who runs those projects?

There is no exact distinction between experts who work for the Language Observatory and who work for the Cyber Census project. At this start-up phase, the team is composed of the following members:

Yoshiki Mikami, Nagaoka University of Technology, Japan	Project leader
Izumi Suzuki, Nagaoka University of Technology, Japan	Project leader's assistant
Venkataraman Narayanan, Excel Solutions, Singapore/India	Internet technology
Yoshihide Chubachi, Keio University, Japan	Cyber Census system
Mitsugu Miura, NEC, Japan	Language Observatory system
Virach Sornlertlamvanich, Communication Research Institute, Thailand	SE Asian languages
Ahmed Zaki abu Bakar, Universiti Teknologi Malaya, Malaysia	Arabic languages
Mitsuru Kano, Nagaoka University of Technology, Japan	South Asian languages
Takashi Yukawa, Nagaoka University of Technology, Japan	Natural language processing
Zavarsky Pavol, Nagaoka University of Technology, Japan	Database system
Hajime Ohiwa, Keio University, Japan	Project adviser
Kazuhiko Machida, Tokyo University of Foreign Studies, Japan	Linguistics adviser
Makoto Minegishi, Tokyo University of Foreign Studies, Japan	Linguistics adviser
Robin Nagano, Miskolc University, Hungary	European languages
Yuko Murayama, Japan Science and Technology Agency	Project Secretary

How those projects actually works?

The Language Observatory

First, crawler robots visit pages on the Internet at least once a year, and fetch text content. Every of these robots are installed into every 120 PCs, and they work in parallel. These robots return back to the same page regularly so as to produce a periodic report. They trace pages using link buttons until the second or third level.

Next, raw text data is stored in the first-level data-storage PCs and is also back-upped to DVDs. In final, the data is transformed to a database style, which is easier to process its information. The database is stored in the second-level data-storage PCs.

The Cyber Census Project

Language Property Identification Module analyses the page content and identifies language property (language, script and character encoding scheme, etc.) of the page. Although automatic identification techniques will be fully employed here, still language experts' contribution should be necessary at this stage. Then the system counts up number of pages according to their language properties.

Our preliminary experiments have revealed that HTML-tag information like “charset”, language is not always reliable, so we need a technology which can identify true language property of pages by analyzing text body itself.

Language and character set identification module

(1) N-gram matching

We developed the shift-codon matching process [3] that is able to respond to either “correct answer” or “unable to detect” where “unable to detect” includes “other than already registered.” On the other hand, conventional methods have no way (or it is difficult for them) to determine whether or not the target is in a particular language, script, and character set already registered.

(2) Word-matching

Almost 80% of the text found on the Internet is written using Latin script and is encoded in ASCII (or ISO/IEC 646, ISO/IEC 8859 series, etc.). But languages of these texts can be European, African or Asian languages (Malay, Indonesian, Pilipino, Turkish, etc.). Major part of ambiguity in the first automatic identification process comes from these Latin texts.

In response to this, an additional automatic identification will be applied to the first level result. At the second level identification process, we employ dictionary-approach to identify language. Short lists of frequently used words, encoded in ASCII, for major Latin script using languages will be used (Word-Matching).

Similar case might happen in other scripts, such as Arabic script (language can be Arabic / Persian / Urdu / etc.), Devanagari script (Hindi / Nepali / Marathi / etc.), and Cyrillic script (Russian / Ukrainian / Mongolian / etc.), then short word lists, encoded in typical character encoding scheme of these scripts, will be created and used for word-matching to complement the first level N-gram matching.

Projects milestones

Work completed so far

- November, 2001 The Cyber Census project was first openly discussed at SEARCC/SRIG-MLC meeting, Auckland, New Zealand
- March, 2002 A preliminary experiment was done by Y. Mikami, I. Suzuki, Y. Chubachi , V. Narayanan and D. Rao
- August, 2002 A pass-breaking technique for language property identification, "Shift-Codon-Matching" is published on ACM/TALIP journal
- September, 2003 The Language Observatory Project was selected as one of JST sponsored program "RISTEX"

Work to be completed

- February, 2004 The Language-Observatory Project will be officially launched.
- The First Language Observatory Workshop will be held in conjunction with 5th International Mother Language Day hosted by UNESCO (under planning).
- February, 2005 The first Cyber Census Report will be published.
- February, 2006 The second Cyber Census Report will be published.

Bibliography

[1] Crystal David, English as a global language, Cambridge University Press, p.54, 1997

[2] Yoshiki Mikami, "Digital Divide" among Languages, Bulletin of Language Science and Humanities, Nagaoka University of Technology, No.14, 2000, pp.83-94

[3] Izumi SUZUKI, Yoshiki MIKAMI, Ario OHSATO, Yoshihide CHUBACHI, A Language and Character Set Determination Method Based on N-gram Statistics, ACM Transactions on Asian Language Information Processing, Vol. 1, No.3, September 2002, pp.270-279.