L2/08-052

Unihan Frequency Data Derived From Wikimedia

John H. Jenkins

28 January 2008

As an experiment, I downloaded the Wikimedia archives for the various East Asian langauges in which their content is availble: Standard Written Chinese, Cantonese, classical Chinese, Japanese, and Korean. I wrote a parser to extract from these archives the actual content of the various pages and do a frequency analysis of the Unihan characters found there.

The results are available at <http://homepage.mac.com/jhjenkins/Unicode/WikimediaFrequency.txt>. This is a text file in ASCII (or UTF-8, if you prefer) with Unix line-endings. Each line consists of a Unicode Scalar Value and frequency data separated by a tab. The frequency data consists of a number of fields separated by spaces, and each field consists of a tag and character count separated by a colon. The tags are zh (Standard Written Chinese), zh-yue (Cantoense), zh-classical (classical Chinese), ja (Japanese) and ko (Korean).

Thus the frequency data for U+4E95 is:

```
ja:14785 ko:162 zh:2387 zh-classical:16 zh-yue:65
```

This means that U+4E95 occurs 14,785 times in the Japanese Wikimedia pages, 162 in the Korean ones, 2387 times in the Standard Written Chinese ones, 16 in the classical Chinese ones, and 65 in the Cantonese ones. From this information it is possible to derive the relative frequencies for various characters.

This data is interesting in that it is derived from a readily-identified source and reflects the way these languages are actually written today. (It's also the only source I've been able to find for Cantonese frequency data.)

It's been suggested that this data be made available to the general

public in the form of a Unicode Technical Note. Before we do so, I should at least figure out how to separate the simplified and traditional Chinese Wikimedia pages from each other so we can get separate counts for each.