The Tibetan OCR Project

Don Stilwell, Leonardo Gribaudo, Lee H. MacDonald, Marvin Moser, Chris J. Fynn, Pierre Robillard, Xavier Franc, Robert Taylor and Robert Chilton of the ACIP, and the Tibetan OCR team (Presented by Marvin Moser.)

ABSTRACT

The Tibetan OCR Project seeks to develop a free production-quality Tibetan Optical Character Recognition system. It was started in early 1998 as an email discussion group among Tibetan scholars, transcribers, and software developers. A web site was created to explain the project, and programmers were invited to join. An ftp archive was created to share calligraphic Tibetan texts and linguistic data files. The calligraphy is the basis for a free Tibetan font under development. The linguistic analysis will be used to produce a syllable grammar/dictionary, format converter, and syllable likelihood estimator for use in Tibetan OCR or in other Tibetan programs. Also, several previously commercial Tibetan software packages for the Macintosh and Windows computers have been made available for free download from their authors web pages.

The project has yielded a grammar of Tibetan syllable components based on a 100 MB ACIP transcription of 1000 classic books. Initial analysis has shown that good categorization (with relatively few exceptions) can be achieved by dividing Tibetan syllables into 3 components: characters before the vowel, the vowel itself, and the following characters. The NASA developed CLIPS expert system was used to parse the ACIP data, verifying the grammar rules for syllable component combination. This small database of syllable component frequencies can then be used to create a fuzzy logic syllable divider/checker, an important step towards an accurate character recognition system.

RESULTS

This paper presents only high level results. A paper describing the calculation methods and results in much more detail, as well as contributions by individual members of the Tibetan OCR group can be downloaded from:

http://www.ghg.net/dstilwell/paper2.PDF (about 1.5 MB).

At the highest level of abstraction in our study, we calculated the frequency of occurrence of distinct syllables. Figure 1 shows a plot of the number of times the most frequent 360 syllables were found in 1,031 Tibetan texts of the Asian Classics Input Project (ACIP) database. As we can show, by integrating the area under the curve, the first 360 most common syllables in Tibetan make up 19.8 million or 87% of the 22.8 million syllables in the database. By the time you go through the top 6,760 most common Tibetan syllables in the database, you have exhausted virtually all of the 22.8 million syllables in this 110 MB database.

We have discovered that virtually all of the syllables in database are accounted for by the top 6,760 most frequent Tibetan syllables plus a small list of about 600 syllable exceptions. Strangely, all of these syllables (except the 600 exceptions) can be built from 3 component parts: an initial part (P1) of one or two Tibetan base glyphs, a vowel (V), and a final part (P2) of one or two Tibetan base glyphs. The initial and final forms (P1 and P2) are composed of one or two consonant glyphs. The vowel is indicated either above or below, or both, of the final consonant in the initial form. The final form can contain vowels too, but there are a relatively small number of these final forms. We treat these P2 vowels as just part of the final form in Tibetan syllables.

Figure 1 - Chart of frequency of entire syllables

By use of our expert system, we were able to determine that there are about 259 important one or two glyph initial forms and about 25 one or two glyph final forms. The number of vowels is even smaller. The 259 most important initial forms account for the first one or two base glyphs in most Tibetan syllables. PA (using ACIP notation) occurs the most frequently and CV occurs the least. This list is not the only one that could be made, but these represent observed frequencies in the 110 MB sample of ACIP texts. Not surprisingly, the same 259 initial forms (P1), 5 major vowels (V), and 25 final forms (P2), supplemented by the 600 exceptions (E) syllables, also account for ALL of the syllables in the Rangjung Yeshe Tibetan English Dictionary. The final forms (P2) are only 25 in number and these too are supplemented by the list of 600 exception syllables. Like the initial forms, the final forms are either one or two base glyphs in length. Figure 3 shows the frequency distribution of these final forms. Please consult the online paper at the URL given above for additional analysis of overall vowel frequency, plus frequency data of P1 plotted against P2 on a per vowel basis. That paper also contains details on the expert system programming. It is hoped that this lexical frequency data will soon be applied to optical character recognition algorithms.

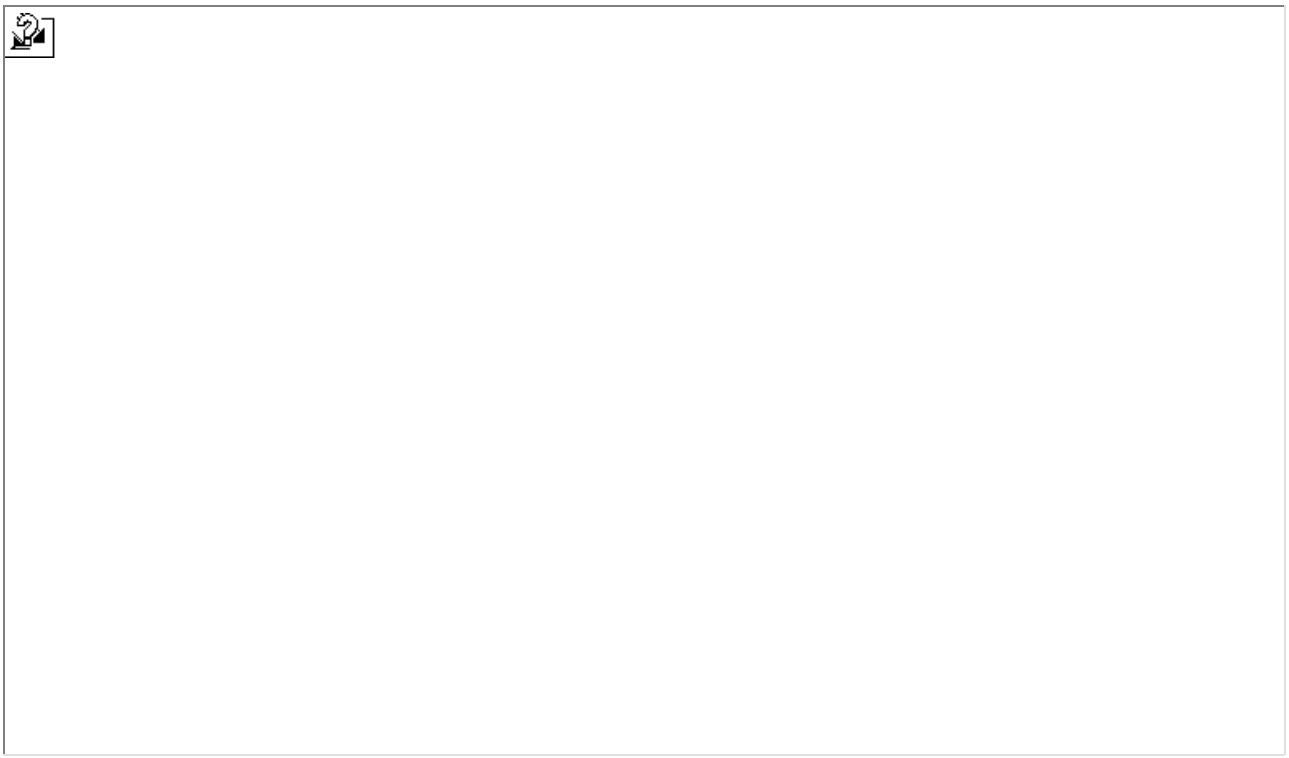

Figure 2 - Chart of frequency of initial syllable parts

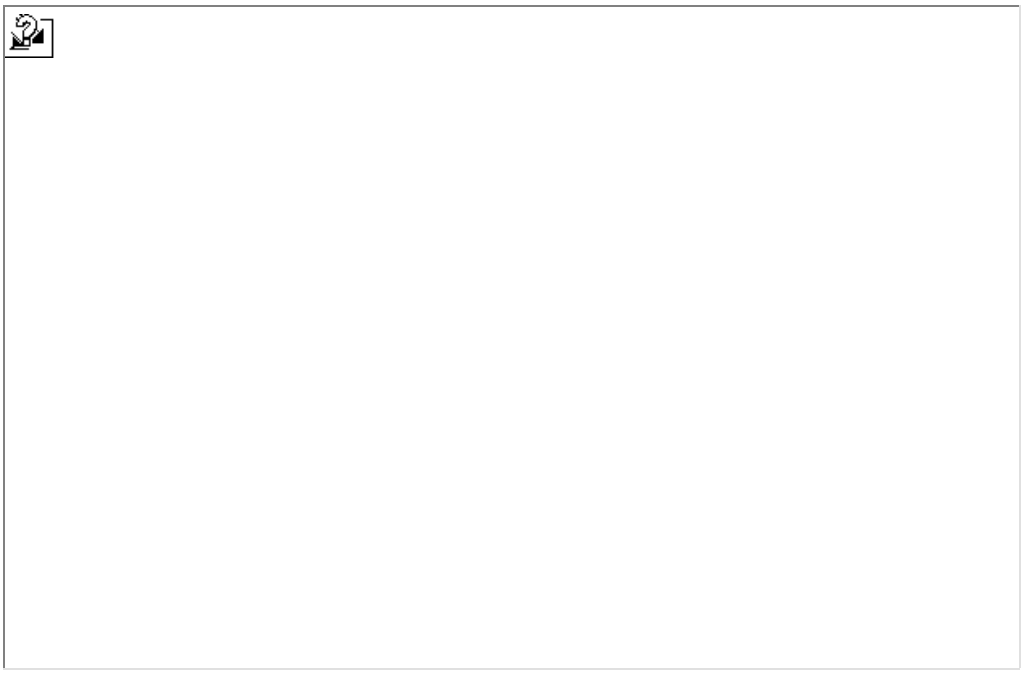

Figure 3 - Chart of frequency of final syllable parts