ABSTRACT

A DESCRIPTION OF

VIETNAMESE IDEOGRAPHIC 𡨸喃 CHỮ NÔM IDEOGRAPHS

Ngô Thanh Nhàn

Linguistic String Project

Computer Science Department – New York University

for Section 2: Cultural Resources

The 5th International Conference on Vietnamese Studies:

Sustainable Development in the Context of Global Change

Vietnam National University, Hanoi

December 15-18, 2016

National Convention Center, Hanoi, Vietnam

Vietnam used a 國語 national script called 𡨸喃 chữ Nôm to represent its language for at least 1,000 years, until the 1920’s. chữ Nôm typically represents Vietnamese dialects at different times in the past. Our collection of known Nôm ideograms is currently at 32,316, covering 20,450 unique ideograms. These are far from being exhaustive. Our repertoire consists of 79,194 ideograms, containing the current UniHan version 8.0.0 of 70,595 ideograms (11,110 Nôm ideograms marked Vsource included), plus over 8,500 recently discovered Nôm ideograms. This means that users regularly face newly-found ideograms in Nôm texts, which require new glyphs and new UniHan codes, before they can be
properly represented.

We are motivated by the fact that scholars who are fluent in Nôm intuitively identify ideograms by their graphic parts, especially when spelling them out loud, never having to count the numbers of strokes. This paper proposes a recursive binary deconstruction of known Nôm ideograms using internal regularities of their graphic representation, to arrive at a set of the smallest meaningful units, which can be called orthographic units. This method improves direct dictionary lookup. It will also allow users to represent often seen misspells and newly-found ideograms.

This approach breaks down a sample of 18,806 ideograms successively into orthographic units, together with their UniHan ideographic description characters, or *idc*s. Thus, for example, ideograms 漕 sòng, 蟾 sùng, 柏 song and 瀤 sùng are composed by the same sequence, an *idc* ⚷, a radical + 崇 sùng, in Polish notation:

\[
\begin{align*}
\text{漕} & \rightarrow \text{⿰氵} + \text{崇}; \\
\text{蟾} & \rightarrow \text{⿰口} + \text{崇}; \\
\text{柏} & \rightarrow \text{⿰木} + \text{崇}; \\
\text{𩦼} & \rightarrow \text{⿰虫} + \text{崇}; \\
\end{align*}
\]

in turn, 崇 sùng is composed by 山 son + 宗 tông;

in turn, 宗 tông is composed by 口 口 miên + 示 kỳ;

in turn, 示 is composed by 二 nhi + 小 tiếu; and

perhaps, in turn, 二 nhi is composed by 一 nhất + 一 nhất

We say 一 nhất, (二 nhi,) 小 tiếu, 口 miên, 山 son, 木 mộc, 口 khẩu, 口 thuy, 虫 trườn, and ｚ nhây are orthographic units that successively form 示 kỳ, 宗 tông, 崇 sùng, 漕 sòng, 蟾 sùng, 柏 song, 嗍 song and 瀤 sùng in specific *idc*s. They also form 呀, 淦, 奠, 标, 崙, 涧, 喃, 峒, 稠, 稠, and 𢬑, 𢬑, 𠒌, 𠒅, 吕, 爻, 品, 宀, 宮, 雫, 出, 容, 宁, 未, 末, 本, 末, 杤, 杹, 林, 棕, 森, 桧, 虫, 蜈, 蟾, 號, ... exponentially.

This approach is capable of regenerate the original ideograms, and all
possible, but yet-to-be-found, ideograms in Nôm texts. For example, 蝶 sùng can be described as
\[
蝶 \rightarrow 虫 + [山 + [宀 + [二 + 小]]],
\]
or produced by, 虫 山 宀 二 小 \rightarrow 蝶.

REFERENCES


A SHORT BIOGRAPHY

Ngô Thanh Nhàn, Ph.D. Linguistics, a scholar of New York University Linguistic String Project (1984-present), a fellow and Adjunct Associate Director of Temple

Website: [http://www.cs.nyu.edu/~nhan](http://www.cs.nyu.edu/~nhan).
Contact by email at [nhan@cs.nyu.edu](mailto:nhan@cs.nyu.edu) or [nhan@temple.edu](mailto:nhan@temple.edu).