ABSTRACT

A DESCRIPTION OF

VIETNAMESE IDEOGRAPHIC 守喃 CHỮ NÔM IDEOGRAMS

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 守喃 *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 漴 *sòng*, 蟲 *sùng*, 標 *song*, 嵏 *song* and 崈 *sùng* are composed by the same sequence, an ***idc*** ⿰, a *radical* + 崇 *sùng*, in Polish notation:

漴 → ⿰ 氵 +崇;   嵏 → ⿰ 口+崇;   標 → ⿰ 木+崇;

蟲 → ⿰ 虫+崇;   崈 → ⿰ 崇+丶;

in turn, 崇 *sùng* is composed by ⿱ 山 *sơn* + 宗 *tông*;

in turn, 宗 *tông* is composed by ⿱ 宀 *miên* + 示 *kỳ*;

in turn, 示 is composed by ⿱ 二 *nhị* + 小 *tiểu*; and

perhaps, in turn, 二 *nhị* is composed by ⿱ 一 *nhất* + 一 *nhất*

We say 一 *nhất*, (二 *nhị*,) 小 *tiểu*, 宀 *miên*, 山 *sơn*, 木 *mộc*, 口 *khẩu*, 氵 *thuỷ*, 虫 *trùng*, and 丶 *nháy* are orthographic units that successively form 示 *kỳ*, 宗 *tông*, 崇 *sùng*, 漴 *sòng*, 蟲 *sùng*, 標 *song*, 嵏 *song* and 崈 *sùng* in specific ***idc***'s. They also form 乐, 沶, 奈, 标, 祢, 漴, 嵏, 崜, 椗, 崇, and 亘, 旦, 三, 峃, 叩, 吕, 屾, 品, 宀, 宫, 屾, 出, 枭, 未, 末, 本, 杰, 杏, 束, 杣, 林, 椗, 森, 楇, 虽, 蚩, 蛛, 査, 蚰, … exponentially.

This approach is capable of regenerate the original ideograms, and all

possible, but yet-to-be-found, ideograms in Nôm texts. For example, 蟓
sùng can be described as

$$蟓 \rightarrow \square\square\ 虫 + [\ \boxminus\ 山 + [\ \boxminus\ 宀 + [\ \boxminus\ 二 + 小\ ]\ ]\ ],$$

or produced by, $\square\square\ 虫\ \boxminus\ 山\ \boxminus\ 宀\ \boxminus\ 二\ 小 \rightarrow 蟓$.

REFERENCES

[1]   Thiều Chửu. 1942. *Hán Việt tự điển* [A Sino-Vietnamese ideogram
dictionary]. Hà Nội: Đuốc Tuệ Publishing House. 818 pp. & Glossary 92 pp.

[2]   Nguyễn Quang Xỉ & Vũ Văn Kính. 1971. 字典𡨸喃 *Tự điển chữ Nôm* [A
Nôm ideogram dictionary]. Sàigòn: Educational Resource Center.

[3]   Viện Ngôn ngữ học. 1976. *Bảng tra chữ Nôm* [A Nôm ideogram glossary].
Hà Nội: The Social Sciences Publishing House. 428 pp.

[4]   康熙字典 *KangXi Dictionary*. 1981. Taiwan: Trung Tân Library.

[5]   Lê Văn Quán. 1981. *Nghiên cứu về chữ Nôm* [A study in Nôm ideograms].
Hanoi: Social Sciences Publishing House. 232 pp.

[6]   C C Hsieh, C T Chang & J K T Huang. 1990. *On the formalization of glyph the
Chinese language*. A contribution to the AFII Meeting in Tokyo. Document No.
ISO/IEC JTC1/SC2/WG2/IRG N292 (February, 1996 in Cupertino).

[7]   J Đỗ, N T Nhàn, N Hoàng. 1992. A proposal for standard Vietnamese character
encodings in a unified text processing framework. *Computer Standards &
Interfaces* 14:3-10.

[8]   Tiêu chuẩn Việt Nam. 1993. TCVN 5712:1993: *Information Technology –
Vietnamese 8-bit standard code character set for information interchange
(VSCII).* Hanoi, Vietnam.

[9]   Tiêu chuẩn Việt Nam. 1993. TCVN 5773: *Information Technology – The Nôm
16-bit character standard code set for information interchange — Chữ Nôm Việt*.

Hanoi. Vietnam. 59 pp.

[10] N T Nhàn. 1994. Some issues in automatic spell checking of Vietnamese written syllables without an associated spelling dictionary. Paper presented at *Tuần lễ Tin học 4 — The Fourth Biennual Technical Conference and Exhibition* (IW'94), Hồ Chí Minh City, Việt Nam. August 2-6, 1994.

[11] Tiêu chuẩn Việt Nam. 1995. TCVN 6056: *Information Technology – The Nôm 16-bit character standard code set for information interchange — Chữ Nôm Hán*. Hanoi. Vietnam. 62 pp.

[12] Ngô Thanh Nhàn. 1998. A review of dictionary indexing and lookup methods for ideographic scripts (http://www.cs.nyu.edu/~nhan/vsic98.pdf), *Hội nghị Việt học lần thứ nhất* [The First Conference of Vietnamese Studies] at the National University of Hanoi, 7-1998.

[13] Ngô Thanh Nhàn. 2001. *Đơn vị chính tả và các đặc điểm của tiếng Việt: chữ quốc ngữ, hệ la-tinh, chữ Nôm, hệ biểu ý và Unicode/ISO IEC 10646*, [Orthographic unit and characteristics of Vietnamese: the romanized quốc ngữ, latin script, chữ Nôm, ideographic script, and Unicode/ISO IEC 10646]. Vietnam Unicode/ISO 10646 Committee, Hanoi.

[14] Fr. Trần Văn Kiệm. 2004. *Giúp đọc Nôm và Hán Việt* [A Guide to Pronunciation of Nôm and Sino-Vietnamese ideograms]. Đà Nẵng Publishing House & The Vietnamese Nôm Preservation Foundation.

[15] Ngô Thanh Giang, Tô Trọng Đức, Ngô Thanh Nhàn & Ngô Trung Việt. 2006. Building a Hán-Nôm character knowledge base and a Nôm basic character component repertoire, paper presented at the *Nôm 2006: The Second International Nôm Conference* at The Thừa Thiên-Huế Center for Information Technology, Huế, June 1-2, 2006.

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

University *Center for Vietnamese Philosophy, Culture & Society* (Philadelphia, 2006-present)*,* and a teacher of *đàn tranh* [16-string zither] at the *Folk Arts – Cultural Treasures Charter School*, Grades 3-8 (Philadelphia, 2009-present), and the *Mekong Traditional Vietnamese Instrumental Ensemble Troupe* (New York, 2013-present).

Website: http://www.cs.nyu.edu/~nhan.

Contact by email at nhan@cs.nyu.edu or nhan@temple.edu.