

LIÊN HIỆP CÁC HỘI KHOA HỌC VÀ KỸ THUẬT TP. HCM
HỘI NGÔN NGỮ HỌC THÀNH PHỐ HỒ CHÍ MINH

KỶ YẾU

Hội thảo khoa học

100 NĂM CHỮ QUỐC NGỮ

TP. HỒ CHÍ MINH, 21/12/2019

**LIÊN HIỆP CÁC HỘI KHOA HỌC VÀ KỸ THUẬT TP.HCM
HỘI NGÔN NGỮ HỌC THÀNH PHỐ HỒ CHÍ MINH**

KỶ YẾU

Hội thảo khoa học

100 NĂM CHỮ QUỐC NGỮ

TP. HỒ CHÍ MINH, 21/12/2019

- 1 1. VÀI TIỂU KHÚC VỀ CHỮ QUỐC NGỮ NHÌN LẠI 100 NĂM
NAY
GS.TS Đinh Văn Đức
- 8 2. PROPOSALS FOR THE REFORM OF QUỐC NGỮ FROM
THE EARLY 19TH CENTURY TO THE PRESENT –
LINGUISTIC PROBLEMS
GS.TS Nguyễn Văn Hiệp, Lê Nam
- 47 3. 100 NĂM CHỮ QUỐC NGỮ ĐI VÀO CUỘC SỐNG
Vu Gia
- 52 4. CHỮ VIẾT TIẾNG VIỆT CUỐI THẾ KỶ XVIII ĐẦU THẾ KỶ
XIX TRÊN CÚ LIỆU QUỐC NGỮ SÁCH SỔ SANG CHÉP CÁC
VIỆC
PGS.TS. Nguyễn Thị Phương Trang
- 64 5. CHUYỂN TỰ TỰ ĐỘNG TỪ CHỮ NÔM SANG CHỮ QUỐC
NGỮ
*PGS.TS Đinh Điền, Nguyễn Thị Kim Phượng, Diệp Gia Hân, Trần
Nguyễn Sơn Thanh*
- 79 6. CHỮ QUỐC NGỮ – TỪ KHỞI SỰ ĐẾN HOÀN THÀNH
PGS.TS. Đặng Ngọc Lệ - TS. Huỳnh Công Tín
- 93 7. BA CUỐN TỪ ĐIỂN MANG TÊN PHILIPPE BÌNH LƯU TRỮ
TẠI VATICAN
TS. Phạm Thị Kiều Ly
- 100 8. TIẾNG VIỆT TỪ THỜI LM DE RHODES KINH LẠY CHA
(PHẦN 5A)
TS. Nguyễn Cung Thông
- 123 9. MỘT SỐ CÁCH GHI PHỤ ÂM ĐẦU TIẾNG VIỆT TRONG
BÚT TÍCH CỦA HỒ CHÍ MINH
TS. Đỗ Thùy Trang
- 133 10. CHUYỆN VIẾT TIẾNG VIỆT
Hoàng Hồng Minh

- 138 11. CHỮ QUỐC NGŨ: GIẢI TỎA NHỮNG THÀNH KIẾN CÙNG
BẮC BỞ NHỮNG NGỘ NHẬN
TS. Đào Tiến Thi
- 152 12. NHỮNG GIÁ TRỊ NGÔN NGỮ VÀ VĂN HÓA CỦA HAI
CUỐN SÁCH DÒNG TÊN: TỪ ĐIỂN VIỆT BÓ LA VÀ SÁCH SỎ
SANG CHÉP CÁC VIỆC
PGS.TS Hoàng Dũng
- 169 13. NGƯỜI BẢO TRỢ CÁC THỪA SAI DÒNG TÊN KHAI SINH
CHỮ QUỐC NGŨ
Lm. Gioan Võ Đình Đệ
- 185 14. NHỮNG TÒN NGHI VÀ NGỘ NHẬN VỀ FRANCISCO DE
PINA VÀ “CHỮ QUỐC NGŨ TIỀN ĐẮC LỘ”
Nguyễn Thanh Quang
- 204 15. VÀI NHẬN XÉT VỀ VAI TRÒ CỦA HỘI TRUYỀN GIÁO TIN
LÀNH PHÚC ÂM LIÊN HIỆP (CHRISTIAN AND MISSIONARY
ALLIANCE) TRONG TRUYỀN BẢ CHỮ QUỐC NGŨ ĐẦU THẾ
KỶ 20 QUA MỘT SỐ CÔNG TRÌNH DỊCH THUẬT KINH
THÁNH
TS. Văn Phú Quang
- 217 16. NHỮNG ĐÓNG GÓP CỦA NHÓM HÀ TIÊN TỬ TUYỆT
VÀO QUÁ TRÌNH PHÁT TRIỂN CHỮ QUỐC NGŨ Ở NAM BỘ
ThS. Lâm Văn Điền
- 230 17. VÀI SUY NIỆM VỀ FRANCISCO DE PINA VÀ NHỮNG
BƯỚC ĐẦU TIÊN TRONG VIỆC HÌNH THÀNH CHỮ QUỐC
NGŨ
GS.TS. Roland Jacques
- 237 18. SỰ HÌNH THÀNH CHỮ VIẾT CỦA CÁC DÂN TỘC VIỆT
NAM THEO MẪU TỰ LATIN
Đào Quốc Trung
- 245 19. BƯỚC ĐẦU NHẬN XÉT VỀ NGÔN NGỮ TIẾP THỊ
THƯƠNG MẠI TRÊN BÁO CHỮ QUỐC NGŨ XƯA
PGS.TS Đinh Kiều Châu

252 20. NHỮNG ĐÓNG GÓP CỦA AN HÀ BÁO VÀO QUÁ TRÌNH
HIỆN ĐẠI HÓA VĂN HỌC QUỐC NGŨ
ThS. Nguyễn Thị Thu Thủy

263 21. VĂN HỌC VIỆT NAM TỪ CUỐI THẾ KỈ XIX ĐẾN 1932 -
BƯỚC CHUYỂN TỪ VĂN HỌC HÁN NÔM SANG VĂN HỌC
QUỐC NGŨ
TS. Nguyễn Thành Thi

277 22. BIẾN ĐỔI NGŨ PHÁP TIẾNG VIỆT CUỐI THẾ KỶ 19,
TRONG TIẾP XÚC VỚI TIẾNG PHÁP: MẢNG NGHIÊN CỨU
LỚN ĐANG BỊ COI NHẸ
Nguyễn Trọng Thành

284 23. NHỮNG NHẬN THỨC CƠ BẢN VỀ CHỮ QUỐC NGŨ
Trần Chút

CHUYỂN TỰ TỰ ĐỘNG TỪ CHỮ NÔM SANG CHỮ QUỐC NGỮ

Đinh Điền, Nguyễn Thị Kim Phượng,
Diệp Gia Hân, Trần Nguyễn Sơn Thanh

1. GIỚI THIỆU

Việc chuyển tự ở đây là sự thay thế chữ viết từ dạng này sang dạng khác trong cùng một ngôn ngữ. Ví dụ: chuyển tự từ hệ chữ Cyrillic của tiếng Nga sang hệ chữ Latin, như: "Путин" thành "Putin". Việc chuyển tự này rất đơn giản vì có sự tương ứng (gần như) 1-1 giữa một mẫu tự trong hệ chữ Cyrillic với một mẫu tự tương ứng trong hệ chữ Latin. Do cả hai hệ chữ viết trên cùng thuộc loại hình chữ viết ghi âm vị (alphabet) nên việc chuyển tự hoàn toàn được thực hiện một cách tự động, nhanh chóng và chính xác bằng cách tra bảng (ánh xạ).

Tuy nhiên, việc chuyển tự từ chữ Nôm sang chữ Quốc ngữ không đơn giản như thế mà ngược lại vô cùng phức tạp do hai hệ chữ không cùng thuộc một loại hình chữ viết. Chữ Nôm thuộc loại hình chữ viết ghi ý (ideographic) còn chữ Quốc ngữ thuộc loại hình chữ viết ghi âm vị theo cách phân chia 6 loại hình chữ viết trên thế giới trong công trình của Rogers H. [1]. Chính vì có sự khác biệt hoàn toàn về hệ chữ viết giữa chữ Nôm (hệ ghi ý, dựa theo Hán tự) và chữ Quốc ngữ (hệ ghi âm vị, dựa theo hệ chữ Latin) nên chúng ta không thể áp dụng cách tra bảng (ánh xạ) vì không có sự tương ứng 1-1 giữa một chữ Nôm với một chữ Quốc ngữ như trong ví dụ chữ tiếng Nga nói trên.

Thực sự, việc chuyển tự tự động từ chữ Nôm sang chữ Quốc ngữ là bài toán rất khó vì chính con người chúng ta khi đọc chữ Nôm cũng phải "vừa đọc vừa đoán" vì cùng một chữ Nôm có thể được "dịch" (chuyển tự) sang nhiều chữ Quốc ngữ khác nhau. Việc suy đoán này phải dùng đến nhiều tri thức cả trong và ngoài ngôn ngữ (extra-linguistic) như văn hóa, lịch sử, địa lý, tiếng Việt cổ, tiếng địa phương, từ chuyên ngành,... Vì vậy, mà từ trước đến nay, việc chuyển tự tự động chữ Nôm sang chữ Quốc ngữ được xem là bài toán bất khả thi. Nhưng gần đây, sự phát triển vượt bậc của lĩnh vực Trí tuệ nhân tạo với các công nghệ Học máy tiên tiến trong ngành Khoa học máy tính đã khiến cho bài toán này có thể thực hiện được dù chưa thể chính xác hoàn toàn. Với công nghệ Học máy, máy có thể tự "học" được cách chọn ("suy

đoán”) chữ Quốc ngữ phù hợp với từng chữ Nôm thông qua ngữ cảnh trong rất nhiều các bản dịch Nôm – Quốc ngữ trước đó của con người. Do đó, nếu chúng ta “dạy” cho máy tính bằng cách cung cấp (đưa vào kho ngữ liệu huấn luyện) cho máy càng nhiều bản dịch Nôm-Quốc ngữ chuẩn, thì máy sẽ càng “thông minh” hơn và cho kết quả chính xác hơn. Ngoài ra, máy cũng có khả năng tự học để hoàn thiện hơn bằng cách rút kinh nghiệm từ các lỗi dịch sai của máy sau khi con người hiệu đính lại những chỗ dịch sai đó. Quá trình này nếu được lặp lại càng nhiều thì máy sẽ càng cho những bản dịch tốt hơn sau này.

Trong bài báo này, chúng tôi sẽ trình bày việc chuyển tự tự động từ chữ Nôm sang chữ Quốc ngữ bằng công nghệ Học máy có bổ sung thêm tri thức từ điển và mô hình ngôn ngữ theo lĩnh vực. Ngoài phần giới thiệu và kết luận, nội dung của bài báo gồm các các phần: tổng quan về chữ Nôm; các nghiên cứu có liên quan; mô hình đề xuất; thực nghiệm và kết quả; hạn chế và hướng phát triển.

2. TỔNG QUAN VỀ CHỮ NÔM

Chữ Nôm do cha ông chúng ta xây dựng dựa trên chất liệu của chữ Hán vì vậy có nhiều nét tương đồng với chữ Hán. Khác với tất cả các hệ chữ ghi âm, chữ Hán là hệ chữ ghi ý và là hệ chữ ghi ý duy nhất hiện đang sử dụng trên thế giới [1]. Với chữ ghi âm, chúng ta có các ký hiệu để ghi lại các vốn âm vị hay vốn âm tiết của một ngôn ngữ. Còn trong chữ Hán ghi ý, có các ký hiệu để ghi lại các ý niệm cơ bản, như là “mặt trời” (bộ “nhật” 日), “mặt trăng” (bộ “nguyệt” 月), “cây cối” (bộ “mộc” 木), “con người” (bộ “nhân” 人), “nước” (bộ “thủy” 水), “tâm hồn” (bộ “tâm” 心),... gọi là các “bộ thủ” (radical), hay còn gọi là “bộ đầu”. Các bộ thủ trên là thành phần cơ bản, thành phần nguyên tố để xây dựng nên các chữ trong tiếng Hán (Hán tự). Hán tự được tạo bằng 6 cách (lục thư), bao gồm [1]:

- Tượng hình (pictograph): 日 (nhật), 月 (nguyệt), 木 (mộc),...
- Chỉ sự/ghi ý (ideograph): 上 (thượng), 一 (nhất), 本 (bản/bồn),...
- Hội ý (compound ideograph): 信 (tín), 林 (lâm), 森 (sâm),...
- Hình thanh (phono-semantic): 妈(ma):yếu tố âm(马/ma/)+yếu tố ý (女[nữ]),...
- Chuyển chú (derivative cognates): 少 (thiếu/thiểu),...

- Giả tá (rebus): 自 (tự) (vay mượn từ nghĩa gốc "cái mũi"),...

Trong 6 dạng chữ đó, chữ tượng hình hiện chỉ còn dưới 4% (vì vậy nếu ta nói chữ tiếng Hoa là chữ tượng hình thì không còn đúng nữa), còn hơn 90% là chữ hình thanh bằng cách ghép một yếu tố ngữ âm (syllable) với yếu tố ngữ nghĩa (morpheme). Vì vậy, ta có thể thấy sự tương đương giữa một Hán tự với một hình tiết (morpho-syllable) trong tiếng Việt. Trong chữ Hán, bộ thủ ghi ý chính là hình vị (vì là thành phần mang nghĩa nhỏ nhất) và đóng vai trò hạt nhân dùng để sắp xếp từ đầu mục cũng như trong việc tra từ điển. Ví dụ: để tra Hán tự “妈 (ma) “mẹ”, ta tra bộ “nữ” (vì “mẹ” là người nữ) và sau đó tìm đến chữ 妈 theo số nét (3 nét).

Nếu như tiếng Việt có khoảng 10.000 âm tiết thuần Việt khác nhau (không tính các tiếng phiên âm, tượng thanh, chữ ghi các tiếng dân tộc,...), thì trong tiếng Hán cũng có khoảng 13.000 Hán tự khác nhau (không tính các Hán tự cổ, các Hán tự phiên âm tiếng nước ngoài, tượng thanh,...). Mỗi Hán tự này có mã Unicode riêng. Mã Unicode của tiếng Hoa hiện nay được xây dựng dựa trên nhiều bộ mã tiếng Hoa (và các Hán tự được tiếng Hàn, tiếng Nhật, tiếng Việt,... vay mượn) như Big5, GB,... và được tổng hợp thành bảng mã Unicode CJK với nhiều phiên bản. Từ phiên bản đầu tiên (1980) với khoảng 13.000 Hán tự đến nay (2018) đã là 80.000 Hán tự [9].

Hầu hết các chữ Nôm của cha ông chúng ta cũng được xây dựng theo nguyên tắc hình thanh nói trên khi mượn 2 yếu tố từ các Hán tự (một yếu tố ghi âm Nôm và một yếu tố ghi ý). Ví dụ: chữ Nôm 𠂔 có nghĩa là số 3 bao gồm Hán tự 巴 có âm Pinyin là /bā/ và âm Nôm là “ba” để ghi âm; và Hán tự 三 có âm Hán-Việt là “tam” (có nghĩa là số 3) để ghi ý. Còn chữ Nôm 𠂔 có nghĩa là “ba, bố, người cha” trong đó phần ghi âm vẫn là Hán tự 巴 và phần ghi ý là bộ 父 có âm Hán-Việt là “phụ” (có nghĩa là “người cha, bố, ba”). Tương tự, chữ Nôm 𠂔 có nghĩa là số 5 bao gồm Hán tự 南 có âm Pinyin là /nán/ và âm Hán-Việt là “nam” để ghi âm và Hán tự 五 có âm Hán-Việt là “ngũ” (có nghĩa là số 5) để ghi ý. Còn chữ Nôm 𠂔 có nghĩa là “năm (tháng)” trong đó phần ghi âm vẫn là Hán tự 南 và phần ghi ý là chữ 年 có âm Hán-Việt là “niên” (có nghĩa là “năm”) [2].

Ngoài cách cấu tạo theo nguyên tắc hình thanh nói trên, còn một số ít chữ Nôm được xây dựng bằng các nguyên tắc khác, như: giả tá, lập, chuyển chú, thêm dấu phụ (diacritics) để phân biệt âm Nôm và âm Hán-Việt [2].

Vì chữ Nôm chủ yếu được xây dựng theo nguyên tắc hình thanh, nên có nhiều trường hợp các chữ Quốc ngữ có âm gần nhau và phạm trù nghĩa cũng gần nhau sẽ có cùng một chữ Nôm. Điều này xảy ra do giữa tiếng Việt và tiếng Hán tuy cùng là ngôn ngữ có thanh điệu nhưng do tiếng Việt của chúng ta có 6 thanh, còn tiếng Hán chỉ có 4 thanh. Ngoài ra, còn có những nguyên nhân khác, như: do phép tạo chữ khác như: lập, giả tá, chuyển chú hay do chuyển âm. Yếu tố biến âm do vùng miền (tiếng địa phương) hay do lịch sử vì âm Hán-Việt du nhập vào nước ta có sự thay đổi theo thời kỳ (âm Hán cổ trước đời Đường và âm Hán từ đời nhà Đường,...),... Chẳng hạn: chữ Nôm 味 có 2 chữ Quốc ngữ là "mùi" (Hán Việt cổ) và "vị" (Hán Việt từ đời nhà Đường)[3]. Chính vì vậy khiến càng có nhiều chữ Nôm đa trị (cùng một chữ Nôm tương đương nhiều chữ Quốc ngữ, nhiều giá trị).

Quan sát từ điển Nôm – Quốc ngữ, ta thấy khoảng một nửa số chữ Nôm đơn trị và một nửa đa trị, ví dụ như trong bảng 1 bên dưới.

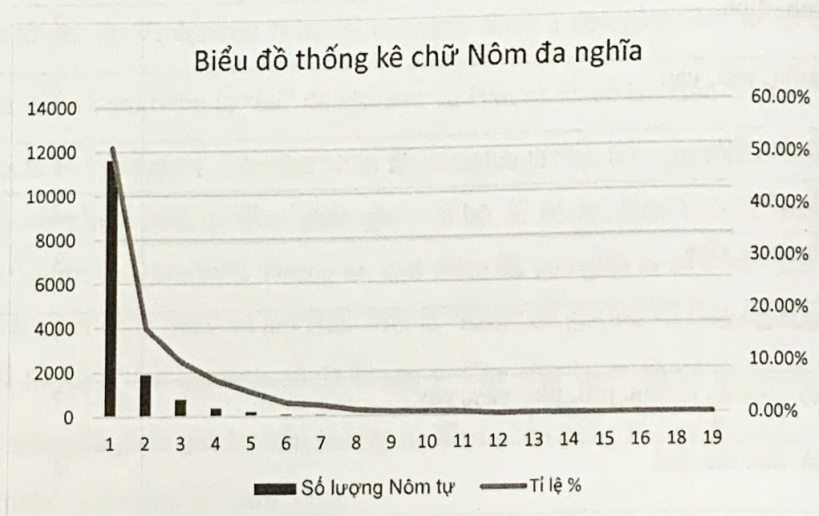
Bảng 1. Ví dụ đa trị của chữ Nôm

Nôm	Quốc ngữ	Số chữ
一	nhất, nhứt	2
丁	đỉnh, đĩnh	2
万	muôn, vàn, vạn	3
与	dữ, dự, đừ	2
丐	cái, gái	2
丑	sấu, xấu, sừu	3
且	thả, vã, vả	3
丕	bậy, chằng, chẳng, phi, phi, vậy, vậy	7
世	thá, thê, thể, thể	4
中	đúng, trong, trung, trúng, truồng	5

Thống kê từ tự điển chữ Nôm [4] gồm 22.264 chữ, ta có kết quả sau:

Bảng 2. Thống kê các chữ Nôm đa trị

Số chữ quốc ngữ tương ứng của một Nôm tự	Số lượng Nôm tự	Tỉ lệ %
1	11.610	52,15%
2	1.907	17,13%
3	787	10,60%
4	384	6,90%
5	209	4,69%
6	94	2,53%
7	66	2,08%
8	30	1,08%
9	20	0,81%
10	16	0,72%
11	11	0,54%
12	3	0,16%
13	3	0,18%
14	2	0,13%
15	1	0,07%
16	1	0,07%
18	1	0,08%
19	1	0,09%



Hình 1. Biểu đồ thống kê chữ Nôm đa trị

Từ kết quả thống kê ở trên, ta thấy gần một nửa (48%) là những chữ Nôm đa trị. Tức là cùng một chữ Nôm nhưng có thể dịch thành nhiều hơn một chữ Quốc ngữ. Cá biệt có trường hợp cùng một chữ Nôm như chữ 折 (mã Unicode 6298h) nhưng tương đương với 19 chữ Quốc ngữ (*chêch, chét, chết, chệt, chiết, chít, chịt, díp, gãy, gảy, giúp, giết, giỡn, nhét, nhít, siết, trét, triếp, xiết*).

Việc chọn chữ Quốc ngữ nào trong số các chữ Quốc ngữ khả dĩ nói trên của cùng một chữ Nôm là vấn đề khó không chỉ đối với máy mà còn đối với cả chính con người. Thậm chí, có những trường hợp rất khó suy đoán như chữ Nôm thứ 3 “𠂔” (mã Unicode 2025Dh) trong câu thứ 12 trong Truyện Kiều của đại thi hào Nguyễn Du: “家資𠂔拱常常壻中” (Gia tư **nghĩ/nghĩ** cũng thường thường bậc trung) có thể chuyển tự thành 2 chữ Quốc ngữ là “nghĩ” (“suy nghĩ”) hoặc “nghì” (“ông ta” trong tiếng Việt cổ) [10]. Với chữ Nôm này, giới học thuật đã tranh cãi từ hơn 50 năm qua với nhiều lập luận, chứng cứ lịch sử, văn học, phương ngữ, mà đến nay vẫn chưa ngã ngũ vì cả hai trường phái đều có cái lý của họ khi suy đoán chữ Nôm đó và cả hai cách dịch đều chấp nhận được (vì đều hợp lý).

Do đó, việc đòi hỏi máy tính cho bản dịch chính xác hoàn toàn là điều không tưởng ít nhất là trong giai đoạn hiện nay và tương lai gần.

3. CÁC CÔNG TRÌNH CÓ LIÊN QUAN

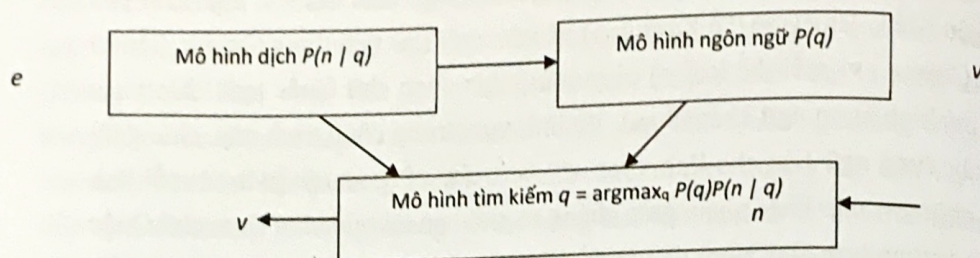
Việc số hóa chữ Nôm đã được một số tác giả (Ngô Thanh Nhân, Nguyễn Quang Hồng, ...) đề xuất và thực hiện từ những năm 1990. Nhờ những đóng góp đó, mà chữ Nôm chúng ta mới có chỗ đứng trong hệ thống bảng mã Unicode của thế giới như ngày hôm nay. Đó chính là tiền đề để thực hiện được các bước tin học hóa sau này, như: lưu trữ, tra cứu, xử lý, chuyển tự tự động,... [11].

Trước hết, phải kể đến phần mềm *Việt Hán Nôm 2002* của Phan Anh Dũng [6] khi cho phép gõ, tra cứu các chữ Hán – Nôm. Phần mềm *Hanosoft* của Tổng Phước Khải [4] có rất nhiều tiện ích liên quan đến việc tra cứu, chuyển tự Hán-Nôm,... Trong các phần mềm này, tác giả đã đưa ra được một công cụ tin học nhằm phiên âm Hán Việt tự động cho các Hán tự và chữ Quốc ngữ cho các chữ Nôm. Tuy nhiên, vấn đề trọng tâm của bài toán là giải quyết các trường hợp đồng tự đa âm vẫn chưa được tác giả giải quyết trong phần mềm đó. Tác giả đành phải chọn ngẫu nhiên một âm Hán Việt hay âm Nôm nào đó trong số các âm Hán Việt hay âm Nôm khả dĩ của nó.

Nếu ta gọi n là câu ngôn ngữ nguồn (chữ Nôm) và q là câu ngôn ngữ đích (chữ Quốc ngữ), ta có phương trình căn bản nhất trong dịch máy thống kê như sau:

$$\hat{q} = \underset{q}{\operatorname{argmax}} P(q) * P(n|q) \quad (1)$$

Qua phương trình (1) thì chúng ta thấy mô hình dịch máy thống kê bao gồm: ước lượng xác suất mô hình ngôn ngữ $P(q)$; ước lượng xác suất mô hình dịch $P(n|q)$; và tìm kiếm một câu chữ Quốc ngữ q tốt nhất sao cho tích $P(q) * P(n|q)$ là lớn nhất. Hình 2 dưới đây thể hiện một hệ dịch thống kê hoàn chỉnh nói trên:



Hình 2. Mô hình chuyển tự

5. THỰC NGHIỆM VÀ KẾT QUẢ

Để thực nghiệm mô hình đề xuất nói trên, chúng tôi sử dụng những ngữ liệu huấn luyện, công cụ huấn luyện, công cụ đánh giá và mô hình ngôn ngữ sau:

5.1 THỰC NGHIỆM

Chúng tôi sử dụng các tự điển và từ điển như Bảng 2 dưới đây:

Bảng 2. Các tự điển/từ điển dùng để huấn luyện

Stt	Mô tả	Kích thước (mục/entry)	Nguồn
1	Tự điển chữ Nôm	913	www.chunom.org
2	Tự điển chữ Nôm	22.264	Hanosoft3
3	Từ điển chữ Nôm	44.485	www.hannom-rcv.org/wi/index.php/標準化
4	Tự điển Hán-Việt	11.285	Hanosoft3
	Tổng cộng:	78.947	
	Sau khi bỏ trùng:	37.358	

Sau đó, chúng tôi dùng chương trình máy tính để gom chung 4 từ điển nói trên rồi loại bỏ các mục từ bị trùng lặp. Về ngữ liệu song ngữ dùng để huấn luyện, chúng tôi sử dụng các tập tin song song Nôm-Quốc ngữ từ website www.chunom.org và Truyện Kiều từ website www.nomfoundation.org (Bảng 3).

Bảng 3. Các ngữ liệu song song Nôm-Quốc ngữ để huấn luyện

Stt	Mô tả	Số câu	Thể loại	Lĩnh vực
1	侗嶺新傳: Cai vàng tân truyện	76	Văn	Văn học
2	經供四時: Kinh cúng tứ thời	27	Văn	Tôn giáo
3	事踢翁狀瓊: Sự tích ông trạng Quỳnh	218	Xuôi	Văn học
4	國音詩集: Quốc âm thi tập	160	Văn	Văn học
5	南無新傳: Nam mô tân truyện	47	Văn	Văn học
6	董天王傳喝 娑姚: Đồng Thiên Vương truyện hát a đào	81	Văn	Văn học
7	佛說大聖末劫真經: Phật thuyết đại thánh mạt kiếp chân kinh	19	Văn	Tôn giáo
8	排歌嶙棋泊: Bài ca răn cờ bạc	108	Văn	Đời sống
9	字喃齣/辭: Chữ Nôm 5/năm	13	Xuôi	Đời sống
10	蛭花新傳: Bướm hoa tân truyện	404	Văn	Văn học
11	蛭花新傳(二): Bướm hoa tân truyện II	404	Văn	Văn học
12	鰲軍新傳: Trống quân tân truyện	79	Văn	Văn học
13	劉平楊禮新傳: Lưu Bình Dương Lễ tân truyện	582	Văn	Văn học
14	訓女演歌: Huấn nữ diễn ca	184	Văn	Văn học
15	潘佩珠小史 (摘) : Phan Bội Châu tiểu sử (trích)	121	Xuôi	Đời sống
16	阮達阮生新傳: Nguyễn Đạt Nguyễn Sinh tân truyện	251	Văn	Đời sống

17	詩陳濟昌: Thơ Trần Tế Xương	36	Văn	Văn học
18	排喝她: Bài hát dĩ	14	Văn	Đời sống
19	褫 終: Lấy chồng chung	8	Văn	Đời sống
20	美女貢胡: Mỹ nữ Công Hồ	189	Văn	Đời sống
21	人問月答: Nhân vấn nguyệt đáp	60	Văn	Văn học
22	桂山三元詩集: 改良賦 Quế sơn tam nguyên thi tập: cải lương phú	153	Văn	Đời sống
23	傳翹: Truyện Kiều	3276	Văn	Văn học
Tổng cộng:		6.259		

Chúng tôi sử dụng công cụ Moses[7] để huấn luyện (training) từ 90% kho ngữ liệu huấn luyện nói trên và dành 10% còn lại để đánh giá (testing).

5.2 KẾT QUẢ

Sau khi huấn luyện, chúng tôi chọn một số câu trong số các câu của tập ngữ liệu đánh giá (nằm trong phần 10% của ngữ liệu dùng để đánh giá, không trùng với 90% phần ngữ liệu dùng để huấn luyện). Điểm BLEU (thông số chuyên dùng để đánh giá kết quả dịch) đạt được là 78,13%. Kết quả từ một số câu (lấy từ những câu đầu và cuối của Truyện Kiều) như Bảng 4 dưới đây.

Bảng 4. Một số ví dụ từ kết quả dịch của hệ thống

Nôm	Bản dịch gốc	Bản dịch hệ thống	Nôm Converter
暮辭 曉 嘆 些	Trăm năm trong cõi người ta	trăm năm 曉 cõi người ta	trăm năm <u>trong</u> cõi người ta
字才字命 害 忒 怙 饒	Chữ tài chữ mệnh khéo là ghét nhau	chữ tài chữ mệnh khéo là ghét nhau	chữ tài chữ mệnh khéo là ghét nhau
歷 戈 叉 局 波 桃	Trải qua một cuộc bể dâu	歷 qua một cuộc bể 桃	歷 qua một cuộc bể 桃
仍 調 離 凭 包 疔 痘 悉	Những điều trông thấy đã đau đón lòng	những điều trông thấy đã đau 痘 lòng	những điều trông thấy đã đau 痘 lòng

追嘆彼喬斯 豐	Lạ gì bĩ sắc tư phong	追 gì bĩ 喬斯豐	追 gì 彼喬斯豐
歪撐帽貝騰 紅打慳	Trời xanh quen với má hồng đánh ghen	trời xanh quen với má hồng đánh ghen	trời xanh quen với má hồng đánh ghen
稿蕙吝換燿 烟	Cảo thơm lần giờ trước đèn	cảo thơm lần 換 trước đèn	稿 thơm lần 換 trước đèn
風情固錄羣 傳史撐	Phong tình có lục còn truyền sử xanh	phong tình có 錄 còn chuyện sử xanh	phong tình có 錄 còn truyện sử xanh
浪辭嘉靖朝 明	Rằng năm Gia Tĩnh triều Minh	rằng năm gia 靖 chiều minh	rằng năm gia 靖 chiều minh
眾方滂湖仁 京凭鑛	Bốn phương phẳng lặng hai kinh vững vàng	bốn phương phẳng láng hai kinh vững vàng	bốn phương phẳng láng hai kinh vững vàng
固茄員外户 王	Có nhà viên ngoại họ Vương	có nhà viên ngoại họ vương	có nhà viên ngoại họ vương
家資偽拱常 常壩中	Gia tư nghĩ cũng thường thường bậc trung	gia tư nghĩ cũng thường thường bậc trung	gia tư nghĩ cũng thường thường bậc trung
巉台閑事嵯歪	Ngẫm thay muôn sự bởi trời	Ngẫm thay muôn sự bởi trời	Ngẫm thay muôn sự bởi trời
歪箕包扒𠵿𠵿 固身	Trời kia đã bắt làm người có thân	Trời kia đã bắt làm người có thân	Trời kia đã bắt làm người có thân
扒風塵沛風塵	Bắt phong trần, phải phong trần	bắt phong trần, phải phong trần	bắt phong trần, phải phong trần
朱清高買特 份清高	Cho thanh cao mới được phần thanh cao	cho thanh cao mới được <u>phần</u> thanh cao	cho thanh cao mới được 份 thanh cao

固兜偏為馭 芾	Có đầu thiên vị người nào	có đầu thiên vị người nào	có đầu thiên vị người nào
字才字命淮 濤奇仁	Chữ tài chữ mệnh đôi dào cả hai	chữ tài chữ <u>mệnh</u> <u>đôi</u> dào cả hai	chữ tài chữ <i>mạng</i> 濤 dào cả hai
固才麻悵之 才	Có tài mà cậ chi tài	có tài mà cậ chi tài	có tài mà cậ chi tài
字才連貝字 灾义韻	Chữ tài liền với chữ tai một vần	chữ tài liền với chữ <u>tai</u> một vần	chữ tài liền với chữ 灾 một vần
抱狴褫業抱 身	Đã mang lấy nghiệp vào thân	đã mang lấy nghiệp vào thân	đã mang lấy nghiệp vào thân
拱脊責客至 斯至賒	Cũng đừng trách lần trời gần trời xa	cũng <u>đừng</u> trách <u>lần</u> trời gần trời xa	cũng 脊 trách <i>lần</i> trời gần trời 賒
善根於在惹 些	Thiện căn ở tại lòng ta	thiện căn ở tại lòng ta	thiện căn ở tại lòng ta
字心箕買朋 匹字才	Chữ tâm kia mới bằng ba chữ tài	chữ tâm kia mới bằng ba chữ tài	chữ tâm kia mới bằng ba chữ tài
啞圭執扣容 賤	Lời quê chấp nhật dong dài	lời quê <u>chấp</u> nhật <u>dong</u> dài	lời quê <i>cháp</i> nhật <i>dung</i> dài
模慍拱特义 菲藏更	Mua vui cũng được một vài trống canh	<u>mua</u> vui cũng được một vài trống canh	mô vui cũng được một vài trống canh

5.3 HẠN CHẾ VÀ HƯỚNG PHÁT TRIỂN

Qua thử nghiệm và kết quả trên, ta thấy hệ thống còn một số hạn chế như:

- Một số chữ Nôm hệ thống chưa nhận diện được: Lỗi này chủ yếu do trong tập ngữ liệu huấn luyện thiếu những cặp câu Nôm-Quốc ngữ mà có chứa các chữ Nôm như thế.

- Chọn chữ Quốc ngữ chưa đúng với chữ Nôm trong một số ngữ cảnh: Đây là vấn đề khó khăn nhất, chỉ có thể cải thiện phần lớn chứ rất khó để khắc phục hoàn toàn.

Để khắc phục các hạn chế nói trên, chúng tôi sẽ tiếp tục thu thập kho ngữ liệu huấn luyện lớn hơn, bao quát hơn và được phân chia theo thể loại, lĩnh vực (thêm lĩnh vực y học, lịch sử,...). Ngoài ra, hệ thống sẽ được tích hợp thêm các tri thức về từ vựng chữ Nôm thông qua các từ điển. Hệ thống cũng sẽ được huấn luyện trên các mô hình học máy mới (như mô hình học sâu – Deep learning).

6. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày việc chuyển tự tự động từ văn bản chữ Nôm sang chữ Quốc ngữ thông qua giải pháp học máy dịch thông kê SMT trong Ngôn ngữ học tính toán. Mô hình đề xuất của chúng tôi so với mô hình chuyển tự hiện có (Nôm Converter) có nhiều cải tiến và đã khắc phục được phần nào các lỗi hiện có, như: chưa nhận diện được một số chữ Nôm, chọn chữ Quốc ngữ chưa đúng theo ngữ cảnh. Trong tương lai, nếu hệ thống này được huấn luyện trên các kho ngữ liệu đầy đủ hơn, bao quát hơn và được huấn luyện bằng các mô hình học máy mạnh hơn, chắc chắn độ chính xác sẽ được nâng cao. Ngoài ra, với cơ chế tự học lỗi thông qua quá trình hậu hiệu đính thủ công, bản dịch của hệ thống sẽ ngày càng hoàn thiện hơn. Hệ thống chuyển tự tự động này sẽ là cầu nối về mặt văn tự giữa chữ Nôm trong quá khứ với chữ Quốc ngữ của chúng ta ở hiện tại. Qua hệ thống này, các kho tàng vô giá về lịch sử, địa lý, y học dân tộc, văn học, ... được cha ông ta trong quá khứ viết bằng chữ Nôm, sẽ được thế hệ trẻ ngày nay khai thác, khám phá bằng chính chữ Quốc ngữ của ngày hôm nay.

TÀI LIỆU THAM KHẢO

1. Rogers, H. (2015). *Writing Systems: A Linguistic Approach*: Blackwell.
2. Nguyễn Tuấn Cường (2012), *Diễn cách cấu trúc chữ Nôm Việt*, NXB ĐHQG Hà Nội.
3. Lê Đình Khẩn (2002), *Từ vựng gốc Hán trong tiếng Việt*, NXB ĐH QG TP. HCM.
4. <https://hanosoft-3-0-hanokey-2010.soft112.com> (truy cập ngày 1/7/2019).
5. Nguyễn Quang Hồng (2006), *Từ điển chữ Nôm* (chủ biên). NXB Giáo Dục.
6. <http://www.hannom.org.vn/detail.asp?param=507&Catid=363> (truy cập ngày 15/6/2006)
7. <http://www.statmt.org/moses/> (truy cập ngày 25/4/2014)
8. Dien Dinh, Kiem Hoang, Eduard Hovy (2003), *BTL: an Hybrid Model in the English – Vietnamese Machine Translation System*, Proceedings of the MT Summit IX, Louisiana, USA, pp. 87-94.
9. Đinh Điền (2019), *Từ điển học tính toán*, NXB ĐHQG HCM.
10. Lê Trung Hoa (2016), "Nghĩ về một số từ khó hiểu trong Truyện Kiều", Kiến thức ngày nay, số 11, ra ngày 20/1/2016.
11. <http://dir.vietnam.online.fr/home/vnChuNom.htm> (truy cập ngày 21/5/2014)