

THU THẬP TỰ ĐỘNG CÁC DỊ THỂ CHỮ HÁN - NÔM ĐỂ CẢI THIỆN CHẤT LƯỢNG CHUYỂN TỰ TỰ ĐỘNG TỪ CHỮ NÔM SANG CHỮ QUỐC NGỮ

VÕ NGỌC TÍN¹ - THÁI HOÀNG LÂM²

TRƯƠNG NHẬT ÁNH³ - ĐINH ĐIỀN⁴

Abstract: Throughout the millennium of formation and development, the Sino-Nom scripts encompass a multitude of variants for the same character. This phenomenon has posed many challenges for automatic transliteration systems from Sino-Nom scripts to Vietnamese National scripts as the training corpus lacks comprehensive coverage of these variant characters. To resolve this issue, our research proposes the use of computer software to automatically collect Sino-Nom variants from reputable websites. The implementation method involves the following steps: collecting automatically Sino-Nom variants from the Internet, building a tool for automated data extraction and applying statistical methods to verify the quality of the collected data repository. The results of our proposed method demonstrate a substantial expansion of the Sino-Nom dictionary with a multitude of new variants, thereby enhancing the accuracy of automatic transliteration from Sino-Nom scripts to national scripts.

Keywords: *Sino-Nom characters, variants, automatic transliteration, dictionary*

1. Giới thiệu

Khi bước vào hành trình khám phá và phiên dịch các văn bản cổ được viết bằng chữ Hán và chữ Nôm của Việt Nam sang chữ Quốc ngữ, chúng tôi nhận ra rõ tầm quan trọng của việc giữ gìn và khai thác di sản văn hóa phong phú này. Chữ Hán là văn tự ngoại lai mà nước ta đã mượn để làm công cụ ghi chép, sáng tác,... trong hơn 2 thiên niên kỉ (được cho là khoảng từ đầu công nguyên đến thế kỉ XX); còn chữ Nôm là loại văn tự được cha ông ta sáng tạo dựa trên chất liệu là chữ Hán. Tuy nhiên, một vấn đề thách thức đang đặt ra trước mắt chúng tôi, đó chính là hiện tượng dị thể của các chữ Hán và chữ Nôm gây khó khăn cho hệ thống chuyển tự tự động. Khái niệm “dị thể” ý chỉ cùng một chữ Hán hay chữ Nôm có thể tồn tại nhiều cách viết không giống nhau. Trong bài báo này, chúng tôi tạm gọi một cách ngắn gọn hai loại chữ Hán và chữ Nôm là chữ Hán - Nôm hay văn tự Hán - Nôm. Vì vậy nói đến dị thể chữ Hán - Nôm là nói đến việc những chữ Hán và chữ Nôm có nhiều cách viết khác nhau cho cùng một chữ.

¹ Cử nhân, Trường ĐH Khoa học Tự nhiên - ĐHQG Thành phố Hồ Chí Minh

² Học viên cao học, Trường ĐH Khoa học Tự nhiên - ĐHQG Thành phố Hồ Chí Minh

³ Sinh viên, Trường ĐH Khoa học Tự nhiên - ĐHQG Thành phố Hồ Chí Minh

⁴ PGS.TS, Trường ĐH Khoa học Tự nhiên - ĐHQG Thành phố Hồ Chí Minh

Quyết tâm vượt qua thách thức này, chúng tôi đã quyết định thực hiện một nghiên cứu để bổ sung các dị thể vào từ điển Hán - Nôm của mình. Để thực hiện điều này, chúng tôi đã đưa ra ý tưởng sử dụng công nghệ và phần mềm máy tính để thu thập tự động các dị thể chữ Hán - Nôm thông qua các trang web uy tín trên Internet. Bằng cách thu thập dữ liệu từ những nguồn đáng tin cậy, chúng tôi hi vọng rằng việc bổ sung lượng lớn dị thể vào từ điển sẽ giúp cải thiện hiệu quả và chính xác của hệ thống chuyên tự tự động.

Đóng góp chính trong nghiên cứu của chúng tôi bao gồm: nghiên cứu vấn đề dị thể trong văn tự Hán - Nôm, thu thập các dị thể từ Internet, thực hiện và đánh giá độ hiệu quả của phương pháp thu thập, bổ sung vào từ điển để nâng cao chất lượng chuyên tự.

2. Văn tự Hán - Nôm và vấn đề dị thể

2.1. Khái quát về văn tự Hán - Nôm

Chữ Hán, hay còn được gọi là chữ Nho, là hệ thống chữ viết bắt nguồn từ Trung Quốc và du nhập vào các nước lân cận, trong đó có Việt Nam, được sử dụng để ghi chép các văn bản trọng yếu như các tài liệu học thuật, văn bản tôn giáo và các văn bản chính trị. Với số lượng lớn từ vựng và ngữ pháp phong phú, chữ Hán đã trở thành nền tảng và nguồn gốc cho sự phát triển ngôn ngữ và văn hóa của người Việt.

Trong khi đó, chữ Nôm là loại văn tự ngữ tố - âm tiết dùng để viết tiếng Việt, do người Việt sáng tạo dựa trên chất liệu chữ Hán với thuyết “Lục thư” của Hứa Thận. Sau khi Ngô Quyền xưng vương năm 939, chấm dứt thời kỳ đô hộ của phong kiến phương Bắc, chữ Nôm bắt đầu xuất hiện và dần trở thành một phần vô cùng quan trọng khi được sử dụng để ghi chép phần lớn các tài liệu thuộc nhiều lĩnh vực khác nhau. Chữ Nôm phát triển mạnh mẽ trong suốt gần một nghìn năm cho đến thế kỉ XIX, phản ánh tri thức, văn hóa và tư tưởng của người Việt.

Tuy nhiên đến thế kỉ XX, do sự phát triển của chữ Quốc ngữ và ảnh hưởng của các cuộc cải cách chữ viết, văn tự Hán - Nôm dần thay thế bởi chữ Quốc ngữ trong hầu hết các lĩnh vực và mất đi vai trò phổ biến trong cuộc sống hàng ngày. Hiện nay, việc bảo tồn và nghiên cứu về văn tự Hán - Nôm vẫn được tiếp tục nhằm tôn vinh và khai thác tài nguyên văn hóa truyền thống của người Việt. Các nghiên cứu này không chỉ giúp hiểu rõ hơn về quá khứ, lịch sử và văn hóa của dân tộc mà còn góp phần nâng cao hiệu suất và độ chính xác trong việc xử lí ngôn ngữ Việt trong thời hiện đại.

2.2. Sự tồn tại của dị thể chữ Hán - Nôm

Trong văn tự Hán - Nôm, sự tồn tại của các dị thể chữ Hán - Nôm là một hiện tượng phổ biến và đa dạng. Dị thể là các biến thể của một chữ, có thể khác nhau về hình dạng, phong

cách viết,... Nguyên nhân gây ra sự đa dạng dị thể có thể do sự phát triển tự nhiên của ngôn ngữ theo thời gian, ảnh hưởng từ vùng miền khác nhau, sự sáng tạo cá nhân của các nhà văn, hoặc sự thay đổi trong quy tắc viết chữ. Đồng thời, hiện tượng dị thể này cũng thể hiện sự đa dạng và phong phú của văn hóa và ngôn ngữ trong quá khứ. Việc nghiên cứu và hiểu về dị thể không chỉ giúp ta tìm hiểu sự biến đổi và tiến hóa của ngôn ngữ, mà còn mang lại những thông tin quý giá về văn hóa, tư tưởng, và tư duy của người Việt xưa.



Hình 1. Tám dị thể của chữ "thể" - 体 theo Từ điển Thi Viện [9]

Việc thu thập và phân loại các dị thể, xây dựng các nguồn dữ liệu đầy đủ và chính xác, cùng với việc phát triển các công cụ và phương pháp xử lý dị thể là những hướng nghiên cứu quan trọng để cải thiện quá trình chuyển tự và khai thác tối đa giá trị của văn tự Hán - Nôm ngày nay.

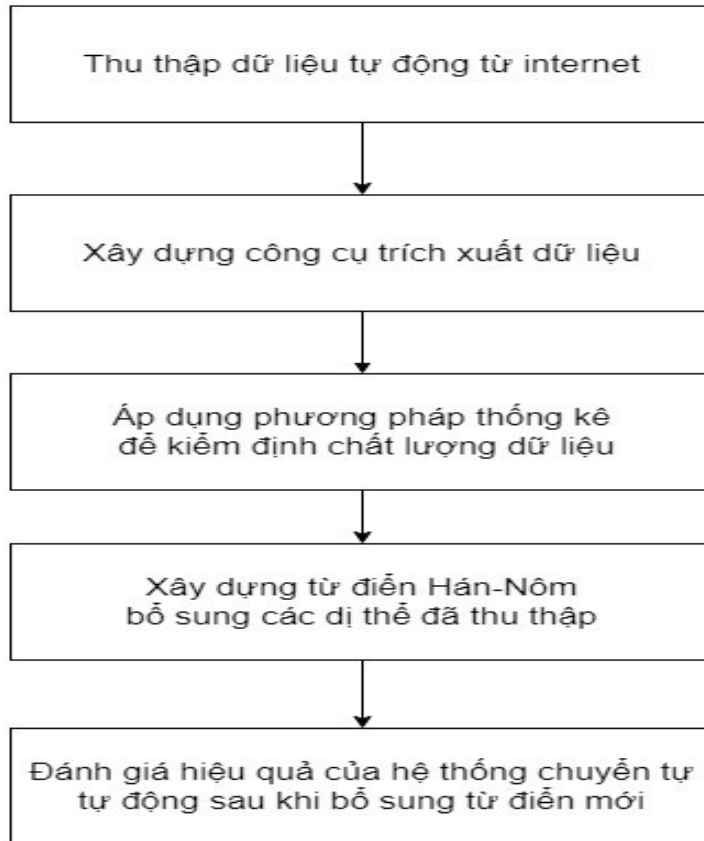
2.3. Vấn đề chuyển tự từ văn tự Hán - Nôm sang chữ Quốc ngữ

Chuyển tự từ văn tự Hán - Nôm sang chữ Quốc ngữ đóng vai trò quan trọng trong việc khám phá và nghiên cứu các văn bản cổ và di sản văn hóa của người Việt Nam. Việc này không chỉ hỗ trợ đáng kể trong công việc dịch thuật, mà còn giúp các tài liệu cổ xưa trở nên dễ dàng tiếp cận đến mọi người ở các trình độ, chuyên môn khác nhau một cách rộng rãi hơn, đặc biệt trong thời đại khi càng có ít người biết và đọc được chữ Nôm.

Tuy nhiên, quá trình chuyển tự đối mặt với nhiều thách thức. Một trong những thách thức quan trọng đó là sự thiếu sót các dị thể trong văn tự Hán - Nôm. Bởi sự đa dạng và khác biệt về hình dạng, phong cách viết và ý nghĩa của chúng, việc không dịch được toàn bộ dị thể có thể gây ra sự hiểu nhầm và làm mất đi thông tin quan trọng trong văn bản gốc. Điều này đặc biệt khó khăn khi văn bản chứa nhiều dị thể hiếm hoặc không phổ biến.

Nhằm đạt được mục tiêu tối ưu hóa chất lượng chuyển đổi từ văn tự Hán - Nôm sang chữ Quốc ngữ, việc mở rộng kho dữ liệu cho các biến thể văn tự là một yếu tố cần thiết để bổ sung và làm phong phú tự/ từ điển Hán - Nôm trong hệ thống dịch. Việc này cho phép chuyển tự tự động văn bản cổ có các chữ Hán - Nôm hiếm gặp chính xác hơn.

3. Phương pháp thu thập tự động các dị thể chữ Hán - Nôm



Hình 2. Sơ đồ minh họa các bước thu thập dữ liệu và đánh giá

3.1. Tự động thu thập dị thể Hán - Nôm từ Internet

Ở mục này, chúng tôi sẽ trình bày quá trình thu thập dữ liệu dị thể Hán - Nôm từ Internet một cách tự động. Để xây dựng một tập dữ liệu đáng tin cậy và đủ lớn, chúng tôi thực hiện các bước sau:

- Xác định nguồn dữ liệu: Chọn các trang web, diễn đàn hoặc tài liệu trực tuyến có chứa thông tin về dị thể Hán - Nôm. Các nguồn này có thể là từ điển trực tuyến, bài viết nghiên cứu có chất lượng và độ tin cậy cao như *Từ điển chữ Nôm dẫn giải* của Nguyễn Quang Hồng [1], *Từ điển Hán Nôm* của Thi viện [2],...

- Thu thập dữ liệu: Sử dụng các kỹ thuật thu thập dữ liệu tự động như “web scraping”, để thu thập các đoạn văn bản chứa dị thể Hán - Nôm từ các nguồn đã xác định và đảm bảo quá trình thu thập được tiến hành một cách chính xác, không vi phạm bản quyền và tuân thủ các quy định pháp luật về bảo vệ dữ liệu cá nhân.

- Tiền xử lí dữ liệu: Dữ liệu thu thập được có thể chứa nhiều và không đồng nhất. Trước khi sử dụng dữ liệu này trong các bước tiếp theo, ta cần thực hiện tiền xử lí như loại bỏ các kí tự không cần thiết, chuyển đổi định dạng phù hợp.

3.2. Xây dựng công cụ trích xuất dữ liệu tự động

Dữ liệu thu thập từ nhiều nguồn trên Internet có các định dạng đa dạng. Để trích xuất các từ dị thể và các chữ gốc tương ứng từ mỗi định dạng, chúng tôi đã viết mã nguồn tùy chỉnh phù hợp. Mỗi định dạng dữ liệu yêu cầu một phương pháp trích xuất riêng biệt để đảm bảo tính chính xác và hiệu quả. Nhờ vào những mã nguồn đa dạng và linh hoạt, chúng tôi đã thành công trong việc trích xuất các từ dị thể Hán - Nôm và các chữ gốc tương ứng và bắt đầu cho việc kiểm tra thủ công.

3.3. Áp dụng phương pháp thống kê để kiểm định chất lượng dữ liệu

Từ dữ liệu tự động đã được trích xuất, chúng tôi đã chọn ngẫu nhiên một mẫu dữ liệu chứa 1500 chữ Nôm, tương ứng với khoảng 10% tổng số lượng chữ đã được thu thập, để tiến hành kiểm định chất lượng.

Bước đầu tiên của phân tích là so sánh số lượng chữ Nôm trong mẫu dữ liệu trích xuất tự động với dữ liệu thực tế đã có sẵn trên Internet. Chúng tôi đã thu thập và đếm số lượng dị thể của chữ Nôm tương ứng từ nguồn đáng tin cậy và tiến hành so sánh với số lượng dị thể mà chúng tôi đã trích xuất. Điều này cho phép chúng tôi đánh giá mức độ đại diện và độ phù hợp của mẫu dữ liệu chữ Nôm được sử dụng trong nghiên cứu.

Tiếp theo, chúng tôi tập trung vào việc kiểm định độ chính xác của tự điển dị thể vừa thu thập. Để làm điều này, chúng tôi đã so sánh thủ công các nét của các chữ đã trích xuất với các hình ảnh chữ Hán - Nôm tương ứng từ nơi thu thập. Đánh giá sự tương đồng và sai khác giữa các nét dị thể giúp chúng tôi đánh giá chất lượng của quá trình trích xuất dữ liệu tự động đã thực hiện bên trên.

Các kết quả từ bước kiểm định được sử dụng để đánh giá hiệu suất và độ tin cậy của phương pháp tự động thu thập dữ liệu dị thể Hán - Nôm. Nếu kết quả kiểm định cho thấy mô hình tự động trích xuất dữ liệu có độ chính xác và độ phủ cao, chúng tôi có thể tự tin sử dụng toàn bộ tập dữ liệu đã thu thập để tiến hành các bước tiếp theo trong nghiên cứu. Trong trường hợp phát hiện các sai sót hoặc hạn chế trong phương pháp tự động, chúng tôi sẽ điều chỉnh và cải thiện công cụ trích xuất dữ liệu để đảm bảo chất lượng dữ liệu tốt nhất có thể.

Tổng hợp kết quả từ phân tích kiểm định sẽ giúp chúng tôi đưa ra nhận xét và kết luận về mức độ tin cậy của bộ tự điển dị thể chữ Hán - Nôm đã thu thập bằng phương pháp tự động, từ đó đảm bảo tính chính xác trong quá trình chuyển đổi từ văn tự Hán - Nôm sang chữ Quốc ngữ khi tích hợp bộ tự điển này.

4. Kết quả

Sau khi thu thập và tiền xử lý dữ liệu, chúng tôi kiểm định thủ công và sau đó dùng phương pháp thống kê để đánh giá.

4.1. Kiểm định giả thuyết

Trên tổng số 15271 chữ thu thập được với 33970 dị thể, chúng tôi tiến hành lấy 1500 mẫu tương ứng với khoảng 10% tổng số để tiến hành kiểm tra thủ công và đánh giá độ chính xác, độ đầy đủ của dữ liệu thu thập được.

Tổng số chữ Hán - Nôm	15.271
Tổng số dị thể	33.970

4.1.1. Kiểm định độ chính xác dữ liệu thu thập

Ở phần này, chúng tôi tiến hành kiểm định độ chính xác dữ liệu thu thập được bằng cách đếm thủ công các dị thể bị sai trong tổng 1.500 mẫu kiểm định. Các bước kiểm định thống kê như sau:

- Đặt giả thuyết kiểm định:

- Giả thuyết H_0 : Dữ liệu thu thập tự động không chính xác với nguồn.
- Giả thuyết H_1 : Dữ liệu thu thập là chính xác.

- Các số liệu kiểm định gồm:

- Độ lớn mẫu: $n = 1500$ (tổng số từ được lấy mẫu)
- Số chữ bị sai: $x = 4$
- Số chữ đúng: $n - x = 1500 - 4 = 1496$
- Tỷ lệ chữ bị sai: $p = \frac{x}{n} = \frac{4}{1500} \approx 0.0027$
- Tỷ lệ chữ đúng: $1 - p = 1 - 0.0027 \approx 0.9973$

- Để kiểm tra giả thuyết, chúng tôi sử dụng mức ý nghĩa $\alpha = 0.05$.

• Giá trị giới hạn tại mức ý nghĩa $\alpha = 0.05$ với bậc tự do $df = n - 1 = 1499$ là $t_{\alpha/2, df} \approx t_{0.025, 1499} \approx 1.962$

- Tiến hành tính toán t-score:

$$t = \frac{p - \mu_0}{\sqrt{\frac{p(1-p)}{n}}}$$

$$t = \frac{0.0027 - 0.05}{\sqrt{\frac{0.0027(1-0.0027)}{1500}}} \approx -42.05$$

Với $t = -42.05$ và $t_{0.025, 1499} \approx 1.962$, ta thấy rằng $t < -t_{0.025, 1499}$, do đó ta bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_1 . Kết quả cho thấy rằng dữ liệu thu thập tự động là chính xác.

4.1.2. Kiểm định độ đầy đủ dữ liệu thu thập

Ở phần này, chúng tôi tiến hành kiểm định độ đầy đủ của dữ liệu thu thập được bằng cách đếm thủ công các dị thể bị thiếu trong tổng 1.500 mẫu kiểm định. Các bước kiểm định thống kê như sau:

- Trong tổng số dị thể của 1.500 mẫu kiểm định, có 742 dị thể bị thiếu. Chúng tôi tính tỉ lệ thiếu dị thể của từng mẫu, tỉ lệ được tính bằng bằng số dị thể thiếu trên tổng số dị thể đúng. Tỉ lệ thiếu dị thể trung bình của 1.500 mẫu là 8.9%.

$$\hat{p} = 0.089$$

- Tiếp theo, chúng tôi tính giá trị z-score tương ứng với mức độ tin cậy 95% (z-score cho khoảng tin cậy 95% là khoảng -1.96 đến 1.96):

$$z \approx 1.96$$

- Tính khoảng tin cậy Wilson:

$$Lowerbound \approx \frac{\hat{p} + \frac{z^2}{2n} - z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \approx 0.074$$

$$Upperbound \approx \frac{\hat{p} + \frac{z^2}{2n} + z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \approx 0.106$$

Vậy tỉ lệ của dị thể bị thiếu là từ 0.074 đến 0.106 (tương đương 7.4% đến 10.6%) với khoảng tin cậy 95%. Từ kết quả trên, chúng tôi nhận định phương pháp thu thập tự động có thể đạt hiệu quả từ 89.4% đến 92.6% việc lấy đầy đủ dị thể từ các nguồn dữ liệu trên Internet.

Đây là một hiệu suất khá tốt tuy nhiên chúng tôi có thể tiếp tục cải tiến phương pháp thu thập để đạt độ đầy đủ cao hơn.

4.2. Bổ sung dữ liệu dị thể vào tự điển Hán - Nôm

Kết quả của việc bổ sung dữ liệu dị thể vào tự điển Hán - Nôm đã đạt được những thành tựu đáng kể. Qua quá trình nghiên cứu, chúng tôi đã thành công trong việc thu thập và bổ sung hơn 33.970 dị thể cho 15.271 chữ Nôm vào tự điển Hán - Nôm. Đây là một con số ấn tượng, cho thấy sự đa dạng và phong phú của văn tự Hán - Nôm. Việc bổ sung các dị thể đã mở rộng phạm vi từ vựng và cung cấp thêm thông tin chi tiết về các biến thể chữ Hán - Nôm. Nhờ vào sự đa dạng này, tự điển Hán - Nôm trở nên đáng tin cậy và hữu ích hơn trong việc hiểu và giải thích các từ ngữ và cấu trúc ngữ pháp trong các văn bản Hán - Nôm.

4.3. Nâng cao chất lượng chuyển tự từ văn tự Hán - Nôm sang chữ Quốc ngữ

Kết quả này đồng thời cải thiện hiệu quả chuyển tự từ văn tự Hán - Nôm sang chữ Quốc ngữ. Với sự bổ sung của hơn 33.970 dị thể, quá trình chuyển tự trở nên chính xác hơn và giảm thiểu sai sót. Điều này góp phần nâng cao chất lượng và độ tin cậy của quá trình chuyển tự, từ đó hỗ trợ việc nghiên cứu, giảng dạy và truyền đạt kiến thức về văn tự Hán - Nôm trong cộng đồng người học và nhà nghiên cứu.

4.5. Cải thiện phương pháp

Sau khi tiến hành phân tích kiểm định chất lượng dữ liệu dị thể Hán - Nôm đã thu thập và trích xuất tự động, chúng tôi nhận thấy rằng việc cải thiện là cần thiết để đảm bảo tính chính xác và đáng tin cậy trong quá trình chuyển tự từ văn tự Hán - Nôm sang chữ Quốc ngữ. Dựa vào các kết quả và nhận xét từ phần kiểm định, chúng tôi đã đề ra những phương hướng cải thiện sau:

- Tối ưu hóa quy trình thu thập dữ liệu: Một số nguồn dữ liệu trên Internet có thể chứa thông tin không chính xác hoặc không đầy đủ. Chúng tôi sẽ tiếp tục tối ưu hóa quy trình thu thập dữ liệu để chọn lọc những nguồn đáng tin cậy hơn và loại bỏ thông tin không chính xác, giúp cải thiện độ tin cậy của dữ liệu thu thập.
- Tăng số lượng mẫu dữ liệu kiểm định: Để đảm bảo tính đại diện cho dữ liệu, chúng tôi dự định tăng số lượng mẫu dữ liệu kiểm định lên. Bằng cách sử dụng một mẫu lớn hơn và đa dạng hơn, chúng tôi sẽ có cái nhìn tổng quát và chính xác hơn về chất lượng dữ liệu trích xuất.
- Tối ưu hóa mô hình trích xuất: Dựa vào các sai sót và khó khăn phát hiện trong quá trình trích xuất dữ liệu, chúng tôi sẽ tối ưu hóa mô hình trích xuất để cải thiện độ chính xác và độ phủ của quá trình. Điều này có thể bao gồm việc sử dụng các phương pháp học máy tiên tiến hơn, điều chỉnh các tham số, và thử nghiệm nhiều mô hình khác nhau.

- Xác minh dữ liệu thực tế: Chúng tôi sẽ thực hiện xác minh dữ liệu trích xuất bằng cách so sánh với dữ liệu chữ Nôm từ các nguồn thực tế và đáng tin cậy khác. Bằng cách so sánh với nhiều nguồn dữ liệu khác nhau, chúng tôi có thể đánh giá độ tin cậy và độ phù hợp của dữ liệu trích xuất một cách toàn diện hơn.

- Liên tục cải tiến và đánh giá: Chúng tôi liên tục cải tiến phương pháp thu thập và trích xuất dữ liệu dị thể Hán - Nôm. Việc này đòi hỏi chúng tôi theo dõi kết quả của việc ứng dụng kho dị thể Hán - Nôm trong tiến trình chuyển tự tự động từ chữ Nôm sang chữ Quốc ngữ, từ đó tinh chỉnh và cải thiện các bước tiếp theo để đạt được kết quả tốt hơn và đáng tin cậy hơn.

Qua việc thực hiện các cải tiến và điều chỉnh như trên, chúng tôi hi vọng sẽ nâng cao hiệu suất và độ tin cậy của công cụ tự động thu thập dữ liệu dị thể Hán - Nôm để bộ tự điển ngày càng phong phú và chuẩn xác hơn. Điều này sẽ giúp tiến trình chuyển tự tự động từ chữ Nôm sang chữ Quốc ngữ sẽ được thực hiện một cách chính xác và đáng tin cậy hơn, đồng thời góp phần hỗ trợ nghiên cứu và học tập chữ Hán, chữ Nôm thông qua tự điển dị thể Hán - Nôm này.

5. Kết luận

Nghiên cứu của chúng tôi đã thành công trong việc thu thập tự động một lượng lớn dị thể Hán - Nôm thông qua phần mềm máy tính và các trang web uy tín trên mạng. Phương pháp trích xuất tự động đã cải thiện hiệu quả chuyển tự từ văn tự Hán - Nôm sang chữ Quốc ngữ. Điều này đóng vai trò quan trọng cho việc bổ sung và nâng cao chất lượng tự điển Hán - Nôm, đồng thời hỗ trợ việc nghiên cứu văn bản cổ và di sản văn hóa của Việt Nam.

Hướng đến tương lai, chúng tôi sẽ tiếp tục nghiên cứu và áp dụng những phương pháp tiên tiến, tối ưu hóa công cụ trích xuất, và mở rộng phạm vi thu thập dữ liệu để tối đa hóa hiệu quả và giá trị của nghiên cứu. Chúng tôi hi vọng rằng công trình này sẽ thúc đẩy sự phát triển và bảo tồn di sản văn hóa của Việt Nam, góp phần lan tỏa kiến thức và nhận rộng ứng dụng trong nhiều lĩnh vực học thuật và xã hội.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Đào Duy Anh, *Chữ Nôm: Nguồn gốc, cấu tạo, diễn biến*, NXB KHXH, 1975.
2. Đinh Điền, *Xử lý ngôn ngữ tự nhiên*, NXB ĐHQG Thành phố Hồ Chí Minh, 2006.
3. Đinh Điền, Trang Minh Chiến, Nguyễn Thị Kim Phụng, Nguyễn Hồng Bửu Long, *Chuyển tự tự động từ chữ Nôm sang chữ Quốc ngữ*, In trong *Hội thảo 100 năm chữ Quốc ngữ*, 2019.

4. Đinh Điền, Trang Minh Chiến, Nguyễn Thị Kim Phụng, Nguyễn Hồng Bửu Long, *Chuyển tự tự động từ chữ Nôm sang chữ Quốc ngữ theo hướng tiếp cận dịch máy neural*, In *Fundamental And Applied IT Research*, 2020.
5. Nguyễn Quang Hồng, *Tự điển chữ Nôm dẫn giải*, Viện Hàn Lâm KHXH Việt Nam, NXB KHXH, 2015.
6. Vũ Văn Kinh, *Đại từ điển chữ Nôm*, NXB Văn nghệ Thành phố Hồ Chí Minh, Trung tâm nghiên cứu Quốc học, 2005.
7. Viện nghiên cứu Hán - Nôm, *Di sản Hán Nôm*, <http://hannom.org.vn/default.asp?CatID=131> (7/2023).
8. Chữ Nôm Việt Nam, *Từ điển Hán Nôm*. <https://chunom.net/Tu-dien-Han-Nom>. (7/2023).
9. Thi Viện, *Từ điển Hán - Nôm*. <https://hvdic.thivien.net/>. (7/2023)

Tiếng Anh

10. Dien Dinh, Phuong Nguyen, and Long H. B. Nguyen, *Transliterating Nôm Scripts into Vietnamese National Scripts using Statistical Machine Translation*”, In *International Journal of Advanced Computer Science and Applications*, 2021.
11. Faulkner, Kenneth G., *The tale of the T-score: review and perspective*. Osteoporosis international, 2005.
12. Lam H. Thai, Long H. B. Nguyen, and Dien Dinh, *Improve the automatic transliteration from Nôm scripts into Vietnamese National scripts by integrating Sino–Vietnamese knowledge into Statistical Machine Translation*, In *Proceedings of the 4th International Conference on Information Technology and Computer Communications*, 2022.
13. *Chữ Nôm Resources*. <https://chunom.org/>. (7/2023)