DOI: https://doi.org/10.59294/HIUJS.KHTT.2026.040

XÂY DỰNG VÀ ỨNG DỤNG KHO NGỮ LIỆU SONG NGỮ VIỆT - TRUNG VỀ TỪ HÁN VIỆT TRONG GIẢNG DẠY

Trương Kỳ Tâm*
Trường Đại học Quốc tế Hồng Bàng

TÓM TẮT
Từ Hán Việt là loại từ vựng đặc thù được hình thành qua quá trình giao lưu lâu dài giữa ngôn ngữ và văn hoá Trung - Việt, có tác dụng chuyển di tích cực rõ rệt, và đã được xem như một công cụ học tập hiệu quả đối với người học tiếng Trung. Tuy nhiên, trong quá trình học tiếng Trung, người Việt Nam cũng thường gặp phải những khó khăn do ảnh hưởng của hiện tượng chuyển di tiêu cực từ từ Hán Việt. Do đó, nghiên cứu này đề xuất xây dựng một kho ngữ liệu song ngữ Việt - Trung chuyên biệt về nhóm từ Hán Việt có chuyển di tích cực. Mục tiêu là hỗ trợ người học nhận diện nhanh hơn những từ Hán Việt có nghĩa tương đồng trong tiếng Trung và nâng cao hiệu quả học tập. Kho ngữ liệu được xây dựng dựa trên "Bảng từ vựng từ Hán Việt chuyển di tích cực" đã được thống kê và xuất bản, sử dụng công cụ căn chỉnh Tmxmall và phần mềm xây dựng kho ngữ liệu LancsBox. Kết quả cho thấy, kho ngữ liệu không chỉ giúp người học nhận diện và đối chiếu từ vựng một cách trực quan mà còn cung cấp tài liệu học tập đa lĩnh vực, từ đó giảm tải nhận thức và nâng cao chất lượng giảng dạy tiếng Trung cho người Việt.

Từ khóa: *Từ Hán Việt, chuyển di ngôn ngữ, kho ngữ liệu song ngữ, giảng dạy tiếng Trung*

CONSTRUCTION AND APPLICATION OF A VIETNAMESE-CHINESE BILINGUAL CORPUS OF SINO-VIETNAMESE VOCABULARY IN LANGUAGE TEACHING

Truong Ky Tam

ABSTRACT
Sino-Vietnamese vocabulary is a special lexical category formed through the long-term interaction between Chinese and Vietnamese language and culture. It exerts a clear positive transfer effect and has been regarded as an effective learning tool for learners of Chinese. However, during the process of learning Chinese, Vietnamese learners also frequently encounter difficulties caused by negative transfer from Sino-Vietnamese words. Therefore, this study proposes constructing a specialized Vietnamese - Chinese bilingual corpus focusing on Sino-Vietnamese items that demonstrate positive transfer, with the aim of helping learners identify Chinese equivalents more quickly and improving learning efficiency. The corpus is built based on the "List of Sino-Vietnamese Words with Positive Transfer," which has already been compiled and published, and employs the Tmxmall alignment tool together with the LancsBox corpus-building software. The results show that the corpus not only helps learners visually recognize and compare vocabulary, but also provides multi-field learning materials, thereby reducing cognitive load and enhancing the quality of Chinese language instruction for Vietnamese learners.

Keywords: *Sino-Vietnamese vocabulary, language transfer, bilingual corpus, Chinese language teaching*

* Tác giả liên hệ: Trương Kỳ Tâm, Email: tamtk@hiu.vn
(Ngày nhận bài: 11/4/2026; Ngày nhận bản sửa: 27/4/2026; Ngày duyệt đăng: 04/5/2026)

1. ĐẶT VẤN ĐỀ

Từ Hán Việt và tiếng Trung có mối quan hệ nguồn gốc lịch sử sâu xa, là một lớp từ vựng gốc Hán được tích lũy và lưu truyền qua hàng nghìn năm trong giao lưu văn hóa ngôn ngữ giữa hai nước Trung Quốc và Việt Nam [1]. Từ khi xuất hiện, từ Hán Việt đã dần dần hòa nhập vào hệ thống từ vựng tiếng Việt cùng với sự phát triển của tiếng Việt qua các giai đoạn khác nhau, và cho đến nay đã trở thành một trong những yếu tố ngôn ngữ quan trọng bậc nhất trong tiếng Việt. Trong tiếng Việt hiện đại, từ Hán Việt có số lượng rất lớn, vẫn được sử dụng rộng rãi và giữ một vị trí cực kỳ quan trọng [2], số lượng của chúng thậm chí còn vượt qua cả từ thuần Việt, chiếm khoảng 60% trong từ vựng tiếng Việt, trong từ vựng sử dụng hàng ngày, nhiều nhà ngôn ngữ học Việt Nam còn cho rằng tỷ lệ này thậm chí cao hơn [3]; trong một phạm vi nhỏ hơn, trung bình trong 500 từ vựng thông dụng tiếng Việt, đã có hơn 240 từ là từ Hán Việt, chiếm 48% từ vựng thông dụng [4]. Với mối quan hệ mật thiết như vậy giữa tiếng Việt và tiếng Trung, đặc điểm của từ Hán Việt đã thể hiện ảnh hưởng rõ rệt trên nhiều phương diện của hai ngôn ngữ, chẳng hạn như ảnh hưởng trong việc dịch thuật và học tập lẫn nhau giữa hai ngôn ngữ.

Trên cơ sở đó, nghiên cứu muốn thông qua phần mềm xây dựng kho ngữ liệu LancsBox, xây dựng kho ngữ liệu song ngữ Trung - Việt về từ Hán Việt để tiến hành phân tích, đối sánh, nhằm cung cấp một kho ngữ liệu phong phú hơn về nhóm từ Hán Việt có tác dụng tích cực đối với việc mở rộng vốn từ vựng tiếng Trung của người Việt, qua đó cũng cung cấp một công cụ tra cứu từ Hán Việt trong hai ngôn ngữ Việt - Trung, nhằm giúp người dạy tiếng Trung có thể hướng dẫn người học thông qua từ Hán Việt để học từ vựng tiếng Trung một cách hiệu quả hơn.

2. KHÁI NIỆM VÀ TỔNG QUAN NGHIÊN CỨU

2.1. Khái niệm và tổng quan nghiên cứu từ Hán Việt

"Từ Hán Việt" hay còn được gọi là "tiếng Hán Việt", là một trong những chủ thể nghiên cứu chính của nghiên cứu này. Đây là một lớp từ vựng đặc biệt được hình thành trong lịch sử giao thoa văn hóa ngôn ngữ hàng ngàn năm giữa Trung Quốc và Việt Nam. Để làm rõ nguồn gốc khái niệm, đặc điểm cũng như ảnh hưởng của từ Hán Việt trong tiếng Việt, nghiên cứu đã tổng kết và phân tích trong quyển chuyên khảo "汉越词研究及实用词汇精选" (Nghiên cứu từ Hán Việt và bảng từ vựng ứng dụng chọn lọc) [5] của chính tác giả.

Khái niệm từ Hán Việt đã được định nghĩa và phát triển trong suốt quá trình nghiên cứu về lớp từ vựng này, và tên gọi của "từ Hán Việt" cũng dần thay đổi theo thời gian. Ban đầu, từ Hán Việt được xem là một loại "từ mượn tiếng Trung". Năm 1912, Henri Maspero lần đầu tiên sử dụng thuật ngữ "Âm Hán Việt" để gọi tên khái niệm này. Năm 1948, học giả Vương Lực đề xuất khái niệm "Tiếng Hán Việt". Năm 1979, Nguyễn Tài Cẩn cho rằng từ Hán Việt là một lớp từ vựng có nguồn gốc Hán, và đề xuất khái niệm "Từ gốc Hán ", được nhiều học giả nghiên cứu về từ Hán Việt đồng tình [6]. Năm 1985, dựa trên cách đọc tiếng Việt của "từ Hán gốc", Nguyễn Thiện Giáp đã đặt tên cho "từ Hán gốc có âm đọc được bằng tiếng Việt" là "từ Hán Việt" [7] và được sử dụng chính thức cho đến ngày nay.

2.2. Khái niệm và tổng quan nghiên cứu về chuyển di ngôn ngữ

"Chuyển di ngôn ngữ" (Language Transfer) là một trong những lý thuyết trọng tâm trong lĩnh vực "Tiếp thu Ngôn ngữ Thứ hai" (Second Language Acquisition, SLA). "Chuyển di ngôn ngữ" là một lý thuyết giải thích về tác động của tiếng mẹ đẻ (ngôn ngữ thứ nhất, L1) đối với người học trong quá trình tiếp thu ngôn ngữ mục tiêu (L2). Tác động này không phải là đơn hướng, mà có thể biểu hiện như một tác động tích cực thúc đẩy việc học (chuyển di tích cực), hoặc một tác động tiêu cực cản trở việc học (chuyển di tiêu cực).

Nghiên cứu về lý thuyết chuyển di ngôn ngữ bắt đầu từ những năm 1950, chịu ảnh hưởng của ngôn ngữ học cấu trúc và tâm lý học hành vi. Nhà ngôn ngữ học người Mỹ Lado trong cuốn "Ngôn ngữ học xuyên văn hóa" đã lần đầu tiên đề xuất giả thuyết phân tích đối chiếu (Contrastive Analysis Hypothesis). Lý thuyết này cho rằng, người học ngôn ngữ thứ hai sẽ vô thức chuyển di các hình thái

ngôn ngữ, ý nghĩa và văn hóa từ tiếng mẹ đẻ sang việc học ngôn ngữ mục tiêu, và sự khác biệt giữa tiếng mẹ đẻ và ngôn ngữ mục tiêu là nguồn gốc chính của khó khăn trong học tập [8]. Ví dụ, người học tiếng Anh thường chuyển di trực tiếp cấu trúc "因为……所以……" của tiếng Hán sang tiếng Anh, dẫn đến lỗi cú pháp "because...so...". Nghiên cứu trong giai đoạn này tập trung chủ yếu vào việc đối chiếu các cấu trúc bề mặt như ngữ âm, từ vựng, cú pháp, nhấn mạnh tác động gây nhiễu của chuyển di tiêu cực từ tiếng mẹ đẻ đối với việc tiếp thu ngôn ngữ [9].

Năm 1998, Pavlenko lần đầu tiên sử dụng thuật ngữ "khái niệm chuyển di" (Conceptual Transfer) [10]. Năm 2000, Jarvis đã trình bày một cách hệ thống về lý thuyết này. Khác với chuyển di ngôn ngữ truyền thống, "khái niệm chuyển di" tập trung vào việc người học chuyển di hệ thống khái niệm từ tiếng mẹ đẻ sang ngôn ngữ mục tiêu như thế nào, liên quan đến sự khác biệt nhận thức sâu sắc về khái niệm từ vựng, khái niệm ngữ pháp và khái niệm văn bản. Sự khác biệt này bắt nguồn từ sự khác biệt xuyên ngôn ngữ trong phương thức lưu trữ khái niệm, chứ không phải chỉ là sự tương ứng từ vựng đơn giản. Jarvis và Pavlenko trong cuốn "Ảnh hưởng Liên ngữ trong Ngôn ngữ và Nhận thức" (Crosslinguistic Influence in Language and Cognition) đã đề xuất rằng chuyển di khái niệm có thể được phân tích từ mười chiều như lĩnh vực sử dụng ngôn ngữ, tính định hướng, cấp độ nhận thức [11]. Ví dụ, người học tiếng Trung thể hiện lợi thế trong việc tiếp thu mệnh đề trạng ngữ chỉ thời gian trong tiếng Anh, phản ánh sự tương đồng về khái niệm biểu đạt thời gian giữa tiếng Trung và tiếng Anh, thể hiện cơ sở nhận thức của chuyển di tích cực.

Trong bối cảnh giảng dạy tiếng Trung như một ngôn ngữ thứ hai, do sự khác biệt đáng kể về loại hình học giữa tiếng Trung và tiếng mẹ đẻ của nhiều người học, đặc biệt là tiếng Việt, tiếng Nhật và tiếng Hàn, trên các phương diện như ngữ âm, từ vựng, ngữ pháp, việc nghiên cứu lý thuyết chuyển di ngôn ngữ có ý nghĩa quan trọng cả về mặt lý luận lẫn thực tiễn. Các nghiên cứu liên quan đến ảnh hưởng của từ Hán Việt đối với việc học tiếng Trung cũng chủ yếu xoay quanh chủ đề tác động chuyển di tiêu cực của chúng đối với người Việt học tiếng Trung. Mặc dù không phủ nhận ảnh hưởng “chuyển di tích cực” mà từ Hán Việt mang lại, nhưng lớp từ vựng này vẫn được coi là một "cái bẫy" đối với người học tiếng Trung trong việc học từ vựng tiếng Trung [12], gây ra không ít khó khăn cho họ trong quá trình học tiếng Trung. Quan điểm này chủ yếu bắt nguồn từ việc một số từ Hán Việt trong tiếng Việt đã xuất hiện hiện tượng chuyển dịch ngữ nghĩa, toàn bộ hoặc một phần ý nghĩa của chúng đã xa rời ý nghĩa trong tiếng Trung, dẫn đến việc người học "vì hiểu sai mà dùng sai" các từ Hán Việt này, gây ra không ít gánh nặng học tập cho bản thân [13]. Mặt khác, nguyên nhân dẫn đến lỗi sai khi sử dụng từ Hán Việt của người học còn nằm ở phương pháp học tập không đúng, thiếu giáo trình được biên soạn dành riêng cho từ Hán Việt và việc thiếu đảm bảo về tính khả dụng của kiến thức từ Hán Việt trên mạng Internet.

Tóm lại, các nghiên cứu hiện nay về ảnh hưởng của từ Hán Việt đối với người Việt học tiếng Trung chỉ giới hạn ở việc phân tích trên bình diện lý thuyết và khảo sát trong phạm vi nhỏ về mặt thực tiễn, chưa thông qua kết quả thực nghiệm để làm rõ trong trường hợp người học sử dụng từ Hán Việt như một công cụ học từ vựng tiếng Trung, cụ thể sẽ chịu ảnh hưởng của những yếu tố nào, làm thế nào để khắc phục ảnh hưởng chuyển di tiêu cực của từ Hán Việt đối với người học, và nâng cao khả năng sử dụng cách thức này để học từ vựng tiếng Trung của người học. Những vấn đề này vẫn cần được chứng minh và giải quyết bằng các kết quả nghiên cứu thực nghiệm sâu hơn.

2.3. Tổng quan nghiên cứu trong và ngoài nước về kho ngữ liệu

Những nghiên cứu về kho ngữ liệu (Corpus) trong và ngoài nước hiện nay đã đạt được nhiều thành tựu đáng ghi nhận. Ở Việt Nam, có thể kể đến các công trình của Đinh Điền về ứng dụng kho ngữ liệu trong giảng dạy ngoại ngữ, chẳng hạn như “Ứng dụng kho ngữ liệu tiếng Việt trong giảng dạy tiếng Việt cho người nước ngoài” (2016), và “Applying Korean-Vietnamese Corpus in teaching Vietnamese for Koreans” (2017). Mặc dù nghiên cứu về kho ngữ liệu trong nước đã được quan tâm nhiều hơn, song phần lớn vẫn tập trung vào lĩnh vực giảng dạy tiếng Anh, tiêu biểu là công trình “Xây dựng kho ngữ liệu du lịch song ngữ Việt - Anh gióng hàng mức câu cho dịch máy” của nhóm tác giả

Nguyễn Tiến Hà, Nguyễn Thị Minh Huyền, Nguyễn Minh Hải.

Đối với kho ngữ liệu tiếng Trung Quốc và việc ứng dụng vào giảng dạy, số lượng nghiên cứu vẫn còn hạn chế, chủ yếu mang tính rời rạc và chưa hình thành một hệ thống hoàn chỉnh. Có thể nêu ví dụ như "Nghiên cứu khảo sát quá trình thụ đắc ngôn ngữ của bổ ngữ xu hướng 来, 去 thông qua kho ngữ liệu" của tác giả Lưu Hớn Vũ. Trong quá trình nghiên cứu kho ngữ liệu Trung - Việt, tác giả này cũng đã xây dựng một kho ngữ liệu song ngữ về từ lóng tiếng Trung và phân tích khả năng ứng dụng của kho ngữ liệu đó trong giảng dạy tiếng Trung [14].

Về phương diện nghiên cứu của học giả nước ngoài, có thể kể đến Camiciottoli (2007) với công trình "Kho ngữ liệu về tọa đàm nghiên cứu thương mại" (Business Studies Lecture Corpus). Tác giả đã xây dựng kho ngữ liệu riêng để tiến hành phân tích và lý giải từ vựng theo các góc độ khẩu ngữ, học thuật, khoa học và nghề nghiệp. Đối với kho ngữ liệu Trung-Việt, phần lớn các nghiên cứu là của các tác giả Trung Quốc, với nội dung tập trung vào phương pháp xây dựng kho ngữ liệu trong lĩnh vực công nghệ thông tin, như "Research on the construction method of Chinese-Vietnamese parallel corpus" [15].

Nhìn chung, chúng ta có thể thấy việc nghiên cứu kho ngữ liệu song ngữ Việt-Trung về từ Hán Việt để ứng dụng vào giảng dạy tại Việt Nam và nước ngoài vẫn còn nhiều hạn chế, dù tác dụng từ Hán Việt đã được phần đông các nhà nghiên cứu công nhận và đóng vai trò khá quan trọng trong quá trình dạy và học từ vựng tiếng Trung tại Việt Nam.

3. MỤC TIÊU, PHƯƠNG PHÁP VÀ CÁCH THỨC TIẾN HÀNH NGHIÊN CỨU

Nghiên cứu kết hợp phương pháp nghiên cứu lý thuyết với phương pháp ứng dụng công nghệ trong xây dựng kho ngữ liệu số. Mục tiêu là xây dựng một kho ngữ liệu song ngữ Việt - Trung về nhóm từ Hán Việt có tác dụng chuyển di tích cực trong giảng dạy và học tập từ vựng tiếng Trung. Quá trình xây dựng kho ngữ liệu được thực hiện qua các bước chính sau:

3.1. Xác định mục tiêu và phạm vi

Mục tiêu chính: Xây dựng một kho ngữ liệu song ngữ Việt-Trung chuyên biệt, tập trung vào việc nhận diện, đối chiếu và cung cấp ngữ liệu mẫu cho nhóm từ Hán Việt có chuyển di tích cực.

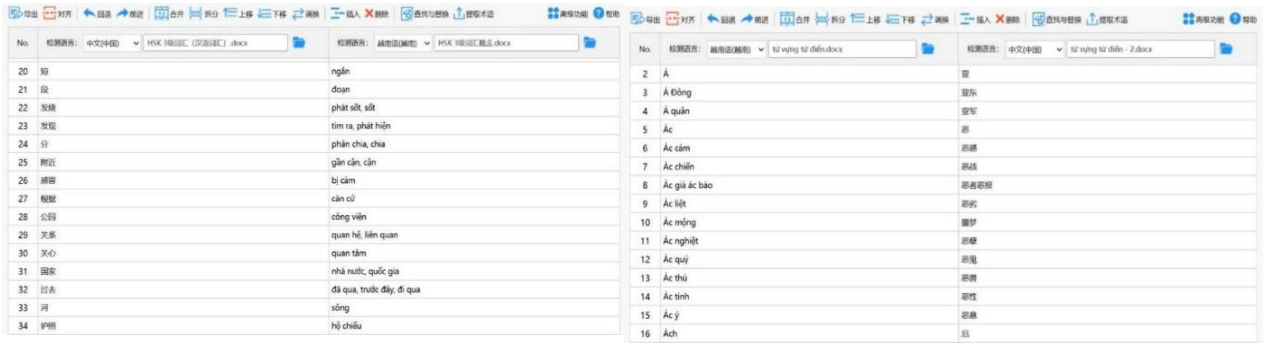
Phạm vi từ vựng: Dựa trên "Bảng từ vựng từ Hán Việt chuyển di tích cực" đã được xác lập trước đó [5], tập trung vào 8 lĩnh vực học thuật và ứng dụng: HSK cấp 3, 4, 5, 6, Kinh tế thương mại cơ bản, Văn hóa xã hội cơ bản, Khoa học kỹ thuật cơ bản, và Nhóm từ vựng từ điển.

Phạm vi ngữ liệu: Thu thập các văn bản, câu, ví dụ có chứa các từ Hán Việt mục tiêu từ các giáo trình tiếng Trung phổ biến như《HSK标准教程》,《汉语教程》,《博雅汉语》,《桥梁》,《发展汉语》và các tài liệu chuyên ngành liên quan.

3.2. Quy trình xây dựng kho ngữ liệu

Để xây dựng kho ngữ liệu song ngữ này, chúng tôi sử dụng phương pháp nghiên cứu của ngành ngôn ngữ học khối liệu (corpus linguistics) nhằm tiến hành căn chỉnh từ vựng Việt - Trung và khái niệm, từ đó xây dựng lên phần ngữ liệu về "Từ vựng đối chiếu", sau đó nhập và tổ chức kho ngữ liệu trên phần mềm LancsBox, và cuối cùng là tích hợp tài liệu học tập đa lĩnh vực. Quá trình xây dựng được chia làm ba giai đoạn chính như trên, tương ứng với cấu trúc của kho ngữ liệu theo các bước cụ thể sau:

(1) Xây dựng phần ngữ liệu về "Từ vựng đối chiếu". Đầu tiên, để chuẩn bị ngữ liệu, chúng tôi tiến hành tách danh sách từ nhóm Hán Việt chuyển di tích cực, từ tiếng Trung tương ứng và khái niệm giải nghĩa thành các file docx riêng biệt. Sau đó sử dụng nền tảng trực tuyến Tmxmall - một công cụ căn chỉnh đa ngữ mạnh mẽ hỗ trợ 46 ngôn ngữ, lần lượt nhập ba cặp dữ liệu cần căn chỉnh: (a) Từ Hán Việt - Từ tiếng Trung, (b) Từ Hán Việt - Khái niệm, (c) Từ tiếng Trung - Khái niệm. Khi công cụ tự động đề xuất căn chỉnh, tiến hành kiểm tra, chỉnh sửa thủ công các căn chỉnh sai lệch thông qua giao diện trực quan của Tmxmall (như Hình 1).



Hình 1. Giai đoạn căn chỉnh từ vựng Việt-Trung và khái niệm qua công cụ Tmxmall

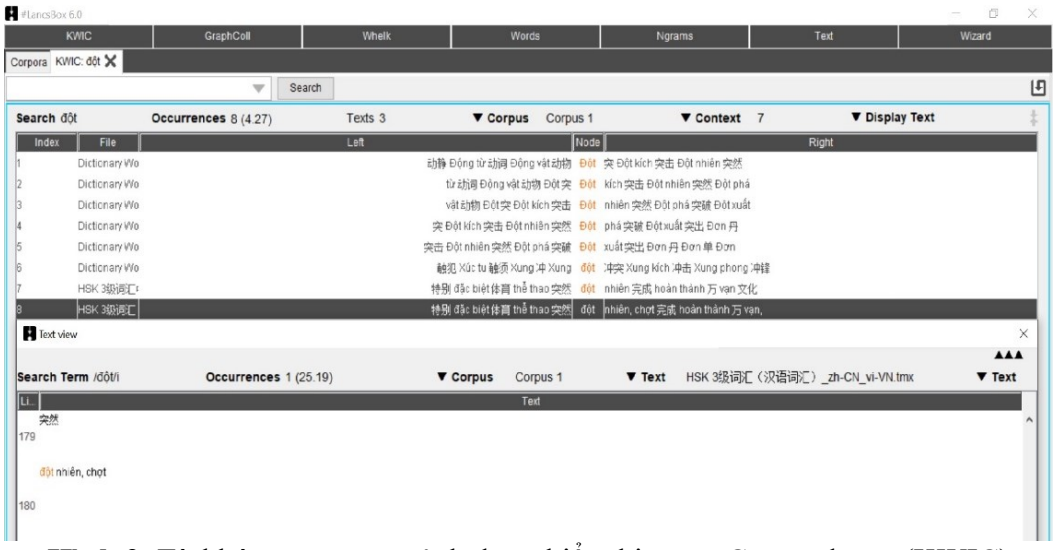
Cuối cùng, xuất toàn bộ kết quả căn chỉnh dưới định dạng chuẩn TMX (Translation Memory eXchange), một định dạng XML phổ biến để trao đổi bộ nhớ dịch như sau:

```
<tmx version="1.4">
  <header creationtool="Tmxmall Aligner" segtype="sentence" adminlang="vi-VN" srclang="vi-VN" datatype="xml" creationdate="20220606T224220Z" creationid="TM STUDIO"/>
  <body>
    <tu creationdate="20220606T224220Z" creationid="TM STUDIO">
      <tuv xml:lang="vi-VN">
        <seg>Trung gian</seg>
      </tuv>
      <tuv xml:lang="zh-CN">
        <seg>中间</seg>
      </tuv>
    <tu creationdate="20220606T224220Z" creationid="TM STUDIO">
      <tuv xml:lang="vi-VN">
        <seg>中间</seg>
      </tuv>
      <tuv xml:lang="zh-CN">
        <seg>ở giữa, trung gian</seg>
      </tuv>
```

(2) Nhập và tổ chức kho ngữ liệu trên phần mềm LancsBox

Sử dụng LancsBox, một phần mềm mã nguồn mở, mạnh mẽ do Đại học Lancaster (Anh) phát triển, được thiết kế cho nghiên cứu ngữ liệu trực quan để nhập file TMX đã xuất từ Tmxmall vào LancsBox. Chức năng cốt lõi của phần mềm này là có thể đọc được ngữ liệu ở nhiều ngôn ngữ và định dạng khác nhau, chẳng hạn như: txt, xml, doc, docx, pdf, csv, zip… đáp ứng nhu cầu xử lý nhiều ngôn ngữ và dạng dữ liệu của các nhà nghiên cứu. Phần mềm còn sở hữu khả năng tra cứu và thống kê mạnh mẽ, hỗ trợ tìm kiếm thông minh và tìm kiếm tinh chỉnh dựa trên từ loại, giúp người dùng nhanh chóng và chính xác hơn trong việc truy xuất thông tin cần thiết.

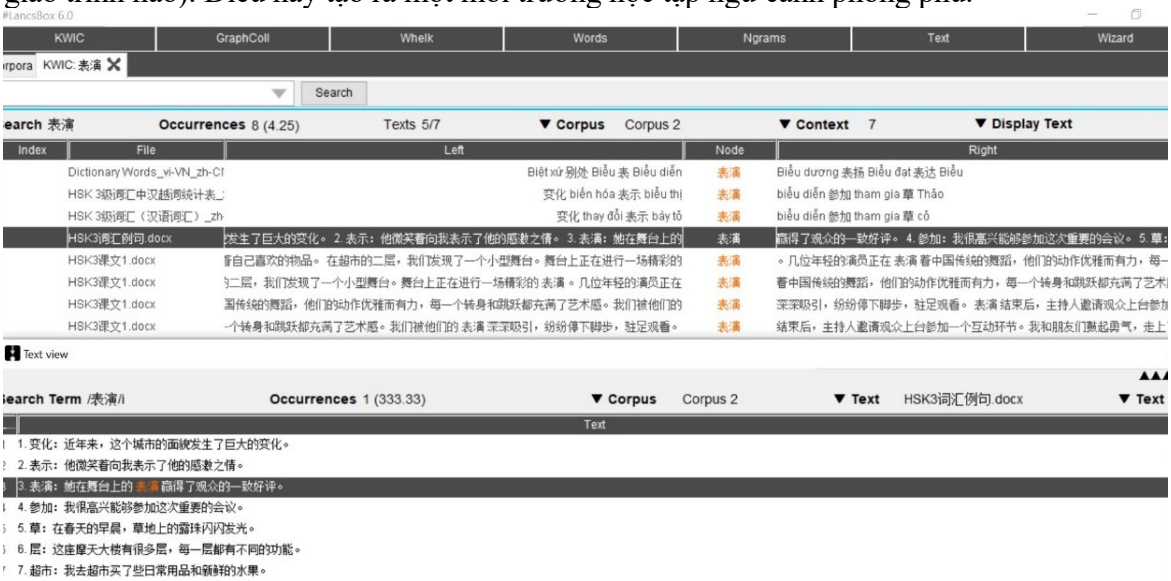
Phần mềm sẽ tự động nhận diện và tổ chức dữ liệu song ngữ. Sau đó sử dụng các tính năng của LancsBox để tổ chức và kiểm tra dữ liệu, như Wordlist để liệt kê tần suất từ vựng; Concordance (KWIC) để hiển thị từ khóa trong ngữ cảnh, cho phép xem các câu/ví dụ chứa từ đó (như Hình 2).



Hình 2. Từ khóa trong ngữ cảnh được hiển thị trong Concordance (KWIC)

(3) Tích hợp tài liệu học tập đa lĩnh vực

Sưu tầm các đoạn văn, bài khóa, ví dụ minh họa từ các giáo trình thuộc 7 lĩnh vực (ngoại trừ lĩnh vực từ điển đã được đưa vào phần khái niệm giải nghĩa của từ vựng) đã nêu là HSK3-4-5-6, Kinh tế cơ bản, Văn hóa cơ bản và Khoa học cơ bản. Sau đó, tiến hành làm sạch văn bản, lưu dưới định dạng .txt hoặc .docx, nhập trực tiếp vào cùng một dự án kho ngữ liệu trong LancsBox. Sau các bước trên, kho ngữ liệu về mặt cấu trúc giờ đây bao gồm hai phần liên thông: Cơ sở dữ liệu từ vựng đã căn chỉnh (Từ Hán Việt <=> Từ tiếng Trung <=> Khái niệm); Kho văn bản mẫu đa lĩnh vực. Khi người dùng tra một từ (ví dụ "表演" - biểu diễn), hệ thống không chỉ hiển thị thông tin đối chiếu mà còn có thể truy vấn để hiển thị tất cả các câu, đoạn văn trong kho văn bản mẫu có chứa từ đó, kèm theo thông tin về nguồn (lĩnh vực nào, giáo trình nào). Điều này tạo ra một môi trường học tập ngữ cảnh phong phú.



Hình 3. Từ khóa "表演-biểu diễn" được hiển thị khi tra cứu trong Concordance (KWIC)

4. KẾT QUẢ VÀ BÀN LUẬN

4.1. Kết quả xây dựng kho ngữ liệu

Sau khi hoàn thành các bước xây dựng kho ngữ liệu, nghiên cứu thành công xây dựng được một "Kho ngữ liệu song ngữ Việt-Trung về từ Hán Việt chuyển di tích cực" với các đặc điểm nổi bật như sau:

(1) Kết quả thống kê "Kho ngữ liệu song ngữ Việt-Trung về từ Hán Việt chuyển di tích cực" (như Bảng 1).

Bảng 1. Kết quả thống kê "Kho ngữ liệu song ngữ Việt-Trung về từ Hán Việt chuyển di tích cực"

STT	Lĩnh vực ngữ liệu	Số lượng từ Hán Việt	Ký hiệu (Tokens)	Phân loại ký hiệu (Types)	Phạm vi thu thập ngữ liệu
1	HSK3	119	418,743	14,892	Giáo trình chuẩn HSK3, Đề thi mẫu, tài liệu ôn tập
2	HSK4	291	851,206	27,615	Giáo trình chuẩn HSK4, Đề thi mẫu, tài liệu ôn tập
3	HSK5	591	1,798,422	51,873	Giáo trình chuẩn HSK5, Đề thi mẫu, tài liệu ôn tập
4	HSK6	772	2,497,831	67,841	Giáo trình chuẩn HSK6, Đề thi mẫu, tài liệu ôn tập
5	Kinh tế cơ bản	407	1,203,569	44,728	Giáo trình tiếng Trung thương mại; Các tài liệu về kinh tế thương mại
6	Văn hóa cơ bản	581	1,498,712	49,632	Giáo trình Văn hóa Trung Quốc; Sách văn hóa, du lịch, lịch sử, di sản văn hóa Trung Quốc
7	Khoa học cơ bản	536	1,297,458	47,916	Giáo trình dự bị đại học tiếng Trung, đề mẫu, tài liệu ôn tập
8	Từ vựng từ điển	8.446	2,998,347	84,572	Từ điển song ngữ, ngữ liệu giải nghĩa
Tổng		11.743	12,564,288	388,069	

(2) Quy mô và cấu trúc

Từ vựng đối chiếu: Kho ngữ liệu chứa tổng cộng 11,743 cặp từ Hán Việt - từ vựng tiếng Trung đã được căn chỉnh chính xác, thuộc 8 lĩnh vực chuyên môn.

Ngữ liệu mẫu: Tích hợp hàng trăm văn bản, câu ví dụ được trích xuất từ các giáo trình uy tín, cung cấp ngữ cảnh sử dụng đa dạng.

Cấu trúc ba lớp: Kho ngữ liệu thiết lập mối quan hệ rõ ràng giữa ba thực thể: Từ Hán Việt <=> Từ tiếng Trung <=> Khái niệm/Nghĩa, tạo thành một mạng lưới tri thức nhỏ.

(3) Tính năng khai thác (thông qua LancsBox)

Tra cứu nhanh: Người dùng có thể tìm kiếm một từ tiếng Việt hoặc tiếng Trung và lập tức nhận được thông tin đối chiếu.

Hiển thị ngữ cảnh (KWIC): Với mỗi từ khóa, hệ thống liệt kê tất cả các câu trong kho ngữ liệu có chứa từ đó, cho phép người học quan sát cách dùng trong thực tế.

Phân tích phối từ (GraphColl): Đây là tính năng đặc biệt giá trị. Ví dụ, khi tra từ "phát triển" (发展), đồ thị có thể cho thấy nó thường đi với các từ như "kinh tế" (经济), "bền vững" (可持续), "khoa học" (科学) trong tiếng Trung, đồng thời cũng hiển thị các collocation tương ứng của "phát triển" trong tiếng Việt. Sự so sánh này giúp người học nắm bắt sắc thái và cách dùng chính xác.

Lọc theo lĩnh vực: Có thể giới hạn kết quả tra cứu chỉ trong các văn bản thuộc lĩnh vực Kinh tế hoặc Văn hóa, phục vụ nhu cầu học tập chuyên sâu.

4.2. Ứng dụng trong giảng dạy và học tập tiếng Trung

"Kho ngữ liệu song ngữ Việt-Trung về từ Hán Việt chuyển di tích cực" sau khi được xây dựng không chỉ là một sản phẩm nghiên cứu, mà còn là một công cụ có giá trị sư phạm trong giảng dạy và học tập tiếng Trung đối với người Việt.

4.2.1. Đối với người học

Kho ngữ liệu đầu tiên có tác dụng "giảm tải nhận thức" cho người học. Thay vì phải ghi nhớ một cách máy móc hoặc suy đoán không chắc chắn, người học có thể tra cứu nhanh để xác nhận liệu một từ Hán Việt có phải là "bạn" (chuyển di tích cực) hay "thù" (chuyển di tiêu cực) của từ tiếng Trung tương ứng. Sự trực quan hóa thông tin qua giao diện phần mềm giúp giảm áp lực ghi nhớ.

Thứ hai, kho ngữ liệu còn có tác dụng giúp cho người học được học từ vựng tiếng Trung qua ngữ cảnh. Việc được tiếp xúc với nhiều câu, đoạn văn mẫu từ các giáo trình thực tế giúp người học hiểu sâu hơn về nghĩa và cách dùng của từ, tránh tình trạng học từ một cách máy móc và không biết cách sử dụng.

Cuối cùng, kho ngữ liệu còn có thể kích thích tính tự giác và nâng cao năng lực tự học của người học. Kho ngữ liệu trở thành một "giáo viên" điện tử luôn sẵn sàng, cho phép người học chủ động khám phá mối liên hệ giữa các từ, mở rộng vốn từ theo chủ đề, tự kiểm tra kiến thức, từ đó nâng cao hứng thú đối với việc học tiếng Trung, đặc biệt là học từ vựng tiếng Trung thông qua từ Hán Việt.

4.2.2. Đối với giáo viên

Kho ngữ liệu có tác dụng hỗ trợ giáo viên tiếng Trung trên các phương diện sau:

Công cụ thiết kế bài giảng: Giáo viên có thể sử dụng kho ngữ liệu để nhanh chóng tìm kiếm và tổng hợp các từ vựng Hán Việt chuyển di tích cực một cách có hệ thống, ví dụ minh họa sinh động cho một chủ đề bài học cụ thể, ví dụ như chủ đề về "Văn hóa". Từ đó giúp người học có thể nắm bắt nghĩa từ vựng nhanh hơn và nâng cao tính hứng thú đối với học tiếng Trung.

Tài liệu thực hành đa dạng: Có thể trích xuất các câu, đoạn văn từ kho ngữ liệu để tạo bài tập điền từ, bài tập dịch, hoặc bài tập nhận diện lỗi sai do chuyển di tiêu cực.

Hỗ trợ giảng giải trực quan: Trong lớp học, giáo viên có thể sử dụng tính năng GraphColl của LancsBox để trình chiếu và giải thích sự khác biệt trong cách kết hợp từ (collocation) giữa tiếng Việt và tiếng Trung, giúp bài giảng sinh động và thuyết phục hơn.

Cá nhân hóa hướng dẫn: Dựa trên nhu cầu của từng học viên (học để thi HSK, học cho chuyên ngành kinh tế...), giáo viên có thể định hướng họ khai thác những phần phù hợp trong kho ngữ liệu.

5. KẾT LUẬN

Nghiên cứu được tiến hành theo cách tiếp cận của ngành ngôn ngữ học khối liệu (corpus linguistics) và chủ yếu sử dụng công cụ căn chỉnh trực tuyến Tmxmall, cùng ứng dụng kho ngữ liệu LancsBox để làm công cụ xây dựng "Kho ngữ liệu song ngữ Việt - Trung về từ Hán Việt chuyển di tích cực". Từ đó cung cấp một khối lượng dữ liệu có cấu trúc để phân tích sâu hơn về hiện tượng chuyển di, biến đổi ngữ nghĩa của từ Hán Việt, cũng như làm công cụ tra cứu về từ Hán Việt chuyển di tích cực dành cho giáo viên và người học tiếng Trung Quốc. Ngoài ra, kho ngữ liệu còn có thể làm tiền đề cho các ứng dụng công nghệ, như: xây dựng từ điển điện tử chuyên biệt, phát triển công cụ kiểm tra lỗi sử dụng từ Hán Việt trong bài viết tiếng Trung của người Việt,v.v. Tuy nhiên, về mặt số lượng từ vựng và ngữ liệu mẫu, kho ngữ liệu cần được tiếp tục bổ sung, cập nhật để ngày càng phong phú và hoàn thiện hơn trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Nguyễn Thiện Giáp, *Từ vựng tiếng Việt*, Hà Nội: Nhà xuất bản Giáo dục Hà Nội, 1985.

[2] 林明华, "汉越词初探", *东南亚研究资料*, no. 04, pp. 111-116, 1986. (Lâm Minh Hoa, "Bước đầu tìm hiểu về từ Hán Việt", *Tư liệu Nghiên cứu Đông Nam Á*,(04):111-116, 1986.)

[3] 阮文康,罗文青, "现代越南语中的汉越词及其变异研究", *广西民族大学学报(哲学社会科学版*, no. 31(04), pp.86-93*)*, 2009 (Nguyễn Văn Khang, La Văn Thanh, "Nghiên cứu về từ Hán Việt và sự biến đổi của chúng trong tiếng Việt hiện đại", *Tạp chí Đại học Dân tộc Quảng Tây (Phiên bản*

Khoa học Xã hội), no. 31(04), pp. 86-93, 2009.)

[4] H. Hua, "论现代越语中的汉越词," 现代外语, no. 03, pp. 38-42, 1990.

[5] Q. Zhang and X. Du, *汉越词研究及实用词汇精选*. Tianjin: 天津大学出版社, 2020.

[6] N. T. Cẩn, *Nguồn gốc và quá trình hình thành ngữ âm Hán Việt*. Hà Nội: Nhà xuất bản Khoa học Xã hội, 1979.

[7] N. T. Giáp, *Từ vựng tiếng Việt*. Hà Nội: Nhà xuất bản Giáo dục, 1985.

[8] R. Lado, *Linguistics Across Cultures*. Ann Arbor: University of Michigan Press, 1957.

[9] T. Odlin, *Language Transfer: Cross-linguistic Influence in Language Learning*. Cambridge: Cambridge University Press, 1989.

[10] A. Pavlenko, *Conceptual Transfer in the Interlingual Lexicon*. Philadelphia: Temple University Press, 1998.

[11] S. Jarvis and A. Pavlenko, *Crosslinguistic Influence in Language and Cognition*. London: Routledge, 2008.

[12] F. L. Ruan, "对越汉语词汇教学中的'陷阱'," *世界汉语教学学会通讯*, no. 2 (总第15期), p. 4, 2012.

[13] P. N. Hàm and L. T. T. Hoài, "Nghiên cứu hiện tượng chuyển dịch ngữ nghĩa của từ gốc Hán trong tiếng Việt," *Tạp chí Ngôn ngữ*, no. 4, pp. 3-9, 2023.

[14] K. T. Trương et al., "Ứng dụng kho ngữ liệu trong giảng dạy từ lóng tiếng Trung Quốc," *Tạp chí Khoa học Trường Đại học Quốc tế Hồng Bàng*, pp. 571-577, 2022.

[15] S. Tu et al., "Research on the construction method of Chinese-Vietnamese parallel corpus," in 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), IEEE, 2019.