VNU Journal of Foreign Studies

Journal homepage: https://jfs.ulis.vnu.edu.vn/

CORPUS LINGUISTICS IN RESEARCH AND TEACHING JAPANESE IN VIETNAM

Tran Thi Minh Phuong*

Faculty of Japanese Language and Culture, VNU University of Languages and International Studies, No. 2 Pham Van Dong, Cau Giay, Hanoi, Vietnam

Received 16 June 2025
Revised 20 October 2025; Accepted 24 November 2025

Abstract: In the context of rapid advancements in science and technology, particularly in computer science, research methods in linguistics have undergone significant innovations. One of the most influential and modern approaches is Corpus Linguistics, which enables the analysis of language based on authentic, objective, and verifiable data. In Vietnam, the application of corpora in foreign language research and teaching, especially in Japanese language education, remains limited. Meanwhile, international studies have demonstrated the effectiveness of corpus-based approaches in various areas such as lexicography, foreign language teaching, translation studies, and discourse analysis. Utilizing corpus data allows learners and researchers to gain a more precise understanding of the grammatical, semantic, and pragmatic features of a language, thereby introducing an empirical and data-driven perspective into applied linguistic research. Based on this practical context, this paper aims to clarify the significance and role of linguistic corpora in Japanese language research and teaching in Vietnam, while surveying and introducing several existing Japanese corpora. Furthermore, the study proposes potential applications of these corpora to support teachers, researchers, and learners in enhancing the effectiveness of Japanese language teaching, learning, and research through real linguistic data.

Keywords: electronic corpus, Japanese language teaching and research, exploitation, application

* Corresponding author.
Email address: phuongttm@vnu.edu.vn
https://doi.org/10.63023/2525-2445/jfs.ulis.5554

NGỮ LIỆU ĐIỆN TỬ TRONG NGHIÊN CỨU VÀ GIẢNG DẠY TIẾNG NHẬT TẠI VIỆT NAM

Trần Thị Minh Phương

Khoa Ngôn ngữ và Văn hóa Nhật Bản, Trường Đại học Ngoại ngữ, Đại học Quốc gia Hà Nội, Số 2 Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Nhận bài ngày 16 tháng 6 năm 2025
Chỉnh sửa ngày 20 tháng 10 năm 2025; Chấp nhận đăng ngày 24 tháng 11 năm 2025

Tóm tắt: Trong bối cảnh khoa học công nghệ phát triển mạnh mẽ, đặc biệt là lĩnh vực khoa học máy tính, các phương pháp nghiên cứu ngôn ngữ cũng có nhiều đổi mới đáng kể. Một trong những hướng tiếp cận hiện đại và có ảnh hưởng sâu rộng là Ngôn ngữ học khối liệu (Corpus Linguistics) - lĩnh vực cho phép phân tích ngôn ngữ dựa trên dữ liệu thực tế, khách quan và có thể kiểm chứng. Tại Việt Nam, việc ứng dụng khối ngữ liệu trong nghiên cứu và giảng dạy ngoại ngữ nói chung, tiếng Nhật nói riêng, vẫn còn hạn chế. Trong khi đó, các nghiên cứu quốc tế đã chứng minh hiệu quả rõ rệt của phương pháp này trong các lĩnh vực như: biên soạn từ điển, giảng dạy ngoại ngữ, nghiên cứu dịch thuật và phân tích diễn ngôn. Việc khai thác ngữ liệu giúp người học và nhà nghiên cứu nắm bắt chính xác hơn đặc điểm ngữ pháp, ngữ nghĩa và ngữ dụng của ngôn ngữ, đồng thời mở ra hướng tiếp cận mới mang tính thực chứng cho các nghiên cứu ngôn ngữ học ứng dụng. Xuất phát từ thực tiễn đó, bài viết này hướng đến việc làm rõ ý nghĩa và vai trò của khối ngữ liệu trong nghiên cứu và giảng dạy tiếng Nhật tại Việt Nam, đồng thời khảo sát, giới thiệu một số nguồn ngữ liệu tiếng Nhật hiện có. Trên cơ sở đó, nghiên cứu đề xuất khả năng ứng dụng phù hợp nhằm hỗ trợ người dạy và người học tiếng Nhật trong việc nâng cao hiệu quả nghiên cứu, giảng dạy và học tập ngôn ngữ dựa trên dữ liệu thực tế.

Từ khóa: khối ngữ liệu điện tử, giảng dạy và nghiên cứu tiếng Nhật, khai thác, ứng dụng

1. Đặt vấn đề

Trong những năm gần đây, các nghiên cứu được thực hiện bởi các thao tác thủ công, dựa trên lý luận và phương pháp phân tích theo trực giác cá nhân đang dần giảm đi do không còn độ tin cậy cao để cho ra các kết quả thuyết phục. Thay vào đó là những nghiên cứu với thao tác tự động, dựa trên các kho tài nguyên ngôn ngữ hay còn gọi là khối ngữ liệu điện tử (e-corpus) đang ngày càng trở nên phổ biến. Từ các khối ngữ liệu điện tử, các học giả trong lĩnh vực nghiên cứu ngôn ngữ có thể khai thác và ứng dụng để phục vụ nhiều mục đích khác nhau như: tìm kiếm, khảo sát, thống kê, đối chiếu để tìm ra các quy luật của ngôn ngữ, quy luật chuyển ngữ, các điểm tương đồng và dị biệt ở các bình diện khác nhau, các cấp độ khác nhau giữa các ngôn ngữ... Những khối ngữ liệu ngày càng trở nên hữu ích cho việc nghiên cứu và giảng dạy ngôn ngữ. Trong tiếng Nhật, đã xuất hiện nhiều khối ngữ liệu ngôn ngữ và khối ngữ liệu người học phục vụ cho việc khai thác ứng dụng vào nghiên cứu và giảng dạy tiếng Nhật. Tuy nhiên, tại Việt Nam, chưa có nhiều người biết đến các khối ngữ liệu điện tử này và đến nay vẫn chưa có bài nghiên cứu nào đi sâu đề cập đến việc khai thác và ứng dụng các khối ngữ liệu điện tử trong việc nghiên cứu và giảng dạy tiếng Nhật. Bài viết này sẽ giới thiệu khái quát về các khối ngữ liệu điện tử hiện có trong tiếng Nhật, đồng thời cũng trình bày các nguồn khai thác và cách thức ứng dụng các khối ngữ liệu điện tử trong việc nghiên cứu và giảng dạy tiếng Nhật cho người Việt Nam.

2. Mục tiêu nghiên cứu

Khối ngữ liệu điện tử tiếng Nhật mang lại nhiều lợi ích quan trọng cho cả nghiên cứu và giảng dạy ngôn ngữ. Thứ nhất, nó cung cấp một nguồn dữ liệu phong phú và đa dạng, giúp người học và nhà nghiên cứu tiếp cận với các mẫu ngôn ngữ thực tế, chính xác và cập nhật. Thứ hai, việc phân tích các dữ liệu trong khối ngữ liệu giúp phát hiện những quy luật ngôn ngữ, xu hướng sử dụng từ vựng và cấu trúc câu trong tiếng Nhật hiện đại. Điều này rất hữu ích trong việc thiết kế giáo trình và phương pháp giảng dạy phù hợp hơn với nhu cầu thực tế của người học. Ngoài ra, khối ngữ liệu cũng hỗ trợ các công việc biên soạn từ điển, nghiên cứu dịch thuật và phát triển các công cụ xử lý ngôn ngữ tự nhiên, góp phần nâng cao chất lượng và hiệu quả của việc học tiếng Nhật. Tuy nhiên, hiện nay ở Việt Nam, các khối ngữ liệu của Nhật cũng chưa được biết đến nhiều và chưa được sử dụng để khai thác và ứng dụng trong nghiên cứu. Xuất phát từ thực trạng này, mục đích của nghiên cứu này sẽ nhằm đưa ra cái nhìn tổng quan về các khối ngữ liệu điện tử hiện có trong tiếng Nhật. Trên cơ sở đó, nghiên cứu đề xuất khả năng có thể ứng dụng và khai thác trong giảng dạy cũng như nghiên cứu tiếng Nhật cho người Việt Nam.

3. Cơ sở lý luận

3.1. Khối ngữ liệu điện tử

Có nhiều định nghĩa khác nhau về "ngữ liệu". Theo Từ điển tiếng Việt của Vietlex Trung tâm từ điển học, NXB Đà Nẵng (2014), "ngữ liệu" là tư liệu ngôn ngữ được dùng làm căn cứ để nghiên cứu ngôn ngữ. Wades and Moje (2000), Göpferich (2006) đã định nghĩa "Ngữ liệu" là một hệ thống tổ chức thống nhất về ngôn ngữ, hoàn chỉnh về nội dung, có chức năng định hướng, do con người tạo ra nhằm sử dụng cho một mục đích xác định. Hay nói cách khác, ngữ liệu là một hình thức giao tiếp bằng lời, bằng văn bản, bằng hệ thống đồ họa để truyền tải ý nghĩa đến người xem.

Dựa vào mục đích và cách xây dựng, McEnery và Wilson (1996) đã chia corpus thành các loại sau: (1) Corpus thô (raw corpus): đơn giản chỉ là tập hợp các dữ liệu mà không có xử lý gì thêm; (2) Corpus được gắn nhãn (tagged corpus): các dữ liệu trong corpus đã được xử lý như phân tích từ, phân tích cú pháp, gắn nhãn từ loại; (3) Parallel Corpus: được sử dụng nhiều trong ứng dụng máy dịch. Ngoài cách chia trên, theo Lê Lâm Thi (2020), corpus cũng có thể được chia theo cấu tạo của nó: (1) Corpus biệt lập (dữ liệu lấy vào 1 cách ngẫu nhiên, biệt lập và không phân biệt với nhau); (2) Corpus theo danh mục (dựa vào các danh mục để chia dữ liệu trong corpus thành các nhóm); (3) Corpus trùng lặp (các dữ liệu trong corpus có thể ở nhiều nhóm cùng lúc); (4) Corpus theo thời gian (các dữ liệu sắp xếp theo thời gian thu thập và thời gian xuất hiện).

Về cách tiếp cận, Lê Tuyết Nga (2020) chỉ ra rằng có hai cách tiếp cận trong ngôn ngữ học khối liệu là: cách tiếp cận dựa vào khối liệu để kiểm chứng lí thuyết (corpus-based) và cách tiếp cận được chỉ dẫn bởi khối liệu để xây dựng lí thuyết (corpus-driven). Hầu hết các học giả trên thế giới đều xác định corpus-based là cách tiếp cận dựa vào khối liệu, có tính diễn dịch, xuất phát từ các giả thuyết, phân tích khối liệu nhằm mục đích kiểm nghiệm, trong khi đó, corpus-driven là cách tiếp cận được chỉ dẫn bởi khối liệu, có tính qui nạp, xuất phát từ dữ liệu và phân tích dữ liệu nhằm mục đích phát hiện, khám phá, từ đó xây dựng luận điểm và lý thuyết.

Về phạm vi ứng dụng, McEneryvà Hardie (2012) chỉ ra rằng khối ngữ liệu đóng vai trò vô cùng quan trọng trong nhiều lĩnh vực liên quan đến ngôn ngữ, cả trong nghiên cứu lý thuyết lẫn ứng dụng công nghệ. Phạm vi ứng dụng khá đa dạng và phong phú. Trong lĩnh vực ngôn

ngữ, các nhà ngôn ngữ học sử dụng khối ngữ liệu để phân tích tần suất từ vựng, cấu trúc câu, hiện tượng ngữ pháp, so sánh ngôn ngữ giữa các vùng miền hoặc theo thời gian, so sánh các đặc trưng ngôn ngữ giữa các thể loại khác nhau như: văn học, báo chí, ngôn ngữ học thuật, hoặc ngôn ngữ mạng xã hội. Trong lĩnh vực giảng dạy ngoại ngữ, giáo viên có thể sử dụng nguồn dữ liệu thực tế để tạo bài tập, nghiên cứu cách dùng từ, xây dựng từ điển và tài liệu giảng dạy hiệu quả. Trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), khối ngữ liệu sẽ là nền tảng để huấn luyện các mô hình máy học trong các ứng dụng như: gắn nhãn từ loại, phân tích cú pháp, tách từ, dịch máy, nhận dạng thực thể, sinh văn bản tự động, v.v.

3.2. Thực trạng xây dựng và nghiên cứu ngữ liệu điện tử ở các nước trên thế giới

Trong thập kỉ vừa qua, nhiều quốc gia đã và đang xúc tiến việc xây dựng các khối liệu trên cơ sở bản ngữ. Trong đó, khối liệu tiếng Anh là nổi bật hơn cả. Các công trình nghiên cứu đã xuất hiện lần đầu tiên vào những những năm 60 thế kỉ XX. Có thể kể đến là: Khối ngữ liệu Brown (Brown University Corpus) chứa khoảng một triệu đơn vị từ và cụm từ sử dụng, được đánh dấu theo dạng hình thái từ; khối ngữ liệu Lancaster/Oslo-Bergen (Lancaster/Oslo-Bergen Corpus (LOB)) - bao gồm khoảng một triệu đơn vị từ và cụm từ sử dụng; Khối ngữ liệu Anh Quốc British National Corpus (BNC) là khối ngữ liệu tiếng Anh có dung lượng lớn nhất hiện nay. Gần đây là sự xuất hiện của Sketch Engine với một bộ ngữ liệu đồ sộ gồm hơn 130 tỉ từ. Tiếp đó, cần kể đến khối liệu tiếng Đức. Theo Lê Tuyết Nga (2020), đây là tập hợp lớn nhất các văn bản bằng tiếng Đức, bao gồm khoảng 2 tỉ đơn vị từ và cụm từ sử dụng. Khối liệu này chứa sơ đồ hình thái cú pháp học dựa trên cơ sở SGML. Hệ thống tự động COSMAS II của khối liệu tiếng Đức cho phép người sử dụng dễ dàng tìm kiếm thống tin chứa trong khối ngữ liệu này. Trung Quốc là nước có khối liệu bản ngữ lớn nhất. Khối liệu tiếng Trung chứa khoảng 1 tỷ đơn vị từ, cụm từ, đang được sử dụng rất rộng rãi và hữu hiệu. Tại Liên bang Nga, ngôn ngữ học khối liệu được bắt đầu nghiên cứu mới chỉ trong vòng hơn thập kỷ trở lại đây nhưng với tốc độ rất nhanh. Hiện nay, ngôn ngữ học khối liệu đang được giảng dạy tại các trường đại học lớn. Việt Nam cũng đang trong quá trình bước đầu xây dựng và nghiên cứu về ngôn ngữ học ngữ liệu. Các công trình nghiên cứu tiêu biểu về "Ngôn ngữ học khối liệu (Corpus) của tác giả Đào Hồng Thu (2007), công trình nghiên cứu về "Ngôn ngữ học máy tính và việc xây dựng từ điển" của hai tác giả Đinh Điền - Hồ Hải Thụy (2011), sách chuyên khảo về Ngôn ngữ học ngữ liệu của tác giả Đinh Điền (2018)…

Tại Nhật Bản, khối ngữ liệu đầu tiên xuất hiện vào năm 1997 của Đại học Kyoto. Đây là khối ngữ liệu chứa khoảng 4 vạn câu được lấy trong các bài đăng trên báo Mainichi phát hành năm 1995. Khối ngữ liệu được dán nhãn các thông tin về hình thái học cũng như đặc điểm về cấu trúc câu. Khối ngữ liệu rất hữu ích cho việc nghiên cứu về quá trình xử lý ngôn ngữ tự nhiên nhưng chi phí sử dụng khá cao nên tỷ lệ khai thác chưa cao. Tiếp đến là khối ngữ liệu khá đồ sộ về văn nói và văn viết của người Nhật bản ngữ được xây dựng bởi Viện nghiên cứu quốc ngữ Nhật Bản. Tuy không được gắn mác là khối ngữ liệu corpus nhưng các thủ pháp thu thập dữ liệu rất gần với phương pháp của ngôn ngữ học khối liệu… Có thể nói, tất cả các khối ngữ liệu điện tử của Nhật hiện nay mặc dù có số lượng dự án xây dựng và khai thác nhiều nhưng thành quả nghiên cứu lại chưa được công bố rộng rãi ra thế giới. Các công trình nghiên cứu định lượng của Viện nghiên cứu quốc ngữ và các tổ chức giáo dục khác ở Nhật Bản không công khai dữ liệu nên cũng là lý do khiến cho nhiều người không biết đến các khối ngữ liệu điện tử của Nhật Bản. Chính vì vậy, bài viết này sẽ tập trung giới thiệu khái quát về các khối ngữ hiện có của tiếng Nhật và đánh giá tiềm năng khai thác cũng như ứng dụng vào việc giảng dạy và nghiên cứu tiếng Nhật cho người Việt Nam.

3. Khai thác và ứng dụng khối ngữ liệu trong giảng dạy và nghiên cứu tiếng Nhật cho người Việt Nam

Có thể khẳng định rằng ngữ liệu điện tử hiện nay đóng vai trò ngày càng quan trọng trong nhiều lĩnh vực, đặc biệt là trong ngôn ngữ học ứng dụng. Việc sử dụng các khối ngữ liệu giúp người học và nhà nghiên cứu tiếp cận dữ liệu ngôn ngữ thực tế, qua đó nâng cao tính chính xác và hiệu quả trong giảng dạy cũng như nghiên cứu. Đối với việc giảng dạy và nghiên cứu tiếng Nhật cho người Việt Nam, các khối ngữ liệu điện tử của Nhật Bản là nguồn tài nguyên quý giá, có thể được khai thác để:

- Phân tích đặc điểm ngữ pháp, từ vựng, và ngữ dụng của tiếng Nhật trong ngữ cảnh thực tế;
- Xây dựng giáo trình, bài tập và tài liệu giảng dạy dựa trên dữ liệu xác thực;
- Nghiên cứu lỗi ngôn ngữ, đối chiếu liên ngữ giữa tiếng Nhật và tiếng Việt;
- Ứng dụng trong công cụ giảng dạy, dịch thuật, và công nghệ ngôn ngữ…

Một số khối ngữ liệu tiêu biểu của Nhật Bản có thể áp dụng trong bối cảnh này bao gồm: BCCWJ (Balanced Corpus of Contemporary Written Japanese), CSJ (Corpus of Spontaneous Japanese), NINJAL-LWP, và Leeds Japanese Concordancer, v.v. Những nguồn này không chỉ cung cấp dữ liệu ngôn ngữ đa dạng mà còn hỗ trợ người học và nhà nghiên cứu Việt Nam trong việc tiếp cận và hiểu sâu hơn về đặc trưng của tiếng Nhật hiện đại. Nội dung và đặc điểm của các khối ngữ liệu được liệt kê ở bảng dưới đây:

Bảng 1

Nội dung và đặc điểm của các khối ngữ liệu trong tiếng Nhật

STT	Tên khối ngữ liệu	Nội dung và đặc điểm	Đơn vị quản lý, vận hành
1	Khối ngữ liệu ngôn ngữ viết cân bằng trong tiếng Nhật hiện đại (Tên tiếng Nhật: 現代日本語書き言葉均衡コーパス (BCCWJ)	* Đây là một khối ngữ liệu được xây dựng bằng tiếng Nhật dùng trong văn viết với quy mô 104,3 triệu từ, được công bố vào năm 2011, gồm 13 thể loại như sách, tạp chí, báo chí, blog, v.v. * Đặc điểm của khối ngữ liệu này là không chỉ bao gồm một thể loại duy nhất, mà được thu thập ngôn ngữ từ nhiều thể loại khác nhau, rất đa dạng phong phú. Dữ liệu được thu thập từ khoảng những năm 1970 đến những năm 2000. Khối ngữ liệu này có thể được tra cứu qua các công cụ trực tuyến trên trang web "https://shonagon.ninjal.ac.jp/ và https://chunagon.ninjal.ac.jp/" * Phần lớn các mẫu được đưa vào trong khối ngữ liệu này được chọn ngẫu nhiên từ tập hợp các dữ liệu xuất bản đã được công bố và dữ liệu sách lưu trữ tại các thư viện công cộng ở khu vực Tokyo. Việc một phần cụ thể nào đó của một cuốn sách hay tạp chí cụ thể được chọn làm mẫu là kết quả hoàn toàn ngẫu nhiên từ quá trình chọn lựa. Không có bất kỳ đánh giá giá trị nào được thực hiện từ góc độ ngôn ngữ học hay văn học. Khối ngữ liệu này có thể được xem là đại diện cho văn viết tiếng Nhật hiện đại, theo	Viện Quốc gia Nghiên cứu Ngôn ngữ Nhật Bản, một tổ chức thuộc Tổ chức Nghiên cứu Văn hóa và Nhân văn Quốc gia (liên kết các trường đại học), và Quỹ Nghiên cứu Khoa học thuộc Bộ Giáo dục, Văn hóa, Thể thao, Khoa học và Công nghệ Nhật Bản tài trợ và quản lý trong khuôn khổ các lĩnh vực nghiên cứu đặc biệt.

		cùng một nghĩa như các cuộc khảo sát dư luận dựa trên phương pháp chọn mẫu ngẫu nhiên do các cơ quan báo chí thực hiện có thể đại diện cho toàn thể người dân Nhật Bản.	
2	Khối ngữ liệu văn nói của người học tiếng Nhật đến từ nhiều quốc gia có tiếng mẹ đẻ khác nhau (Tên tiếng Nhật: 多言語母語の日本語学習者横断コーパス)	Đây là một khối ngữ liệu được thu thập thông qua khảo sát đồng thời cả văn nói và văn viết của 1.000 người học tiếng Nhật đến từ 20 quốc gia và khu vực, với 12 ngôn ngữ mẹ đẻ khác nhau (bao gồm cả Nhật Bản). Những người học tham gia đã được kiểm tra năng lực tiếng Nhật để phân định trình độ. Nhờ vậy, dữ liệu có thể được so sánh theo trình độ, đặc điểm của ngôn ngữ mẹ đẻ, nhiệm vụ thực hiện và môi trường học tập. Hiện tại, khối ngữ liệu này đang trong quá trình xây dựng và tính đến cuối tháng 7 năm 2018, tổng số từ đã thu thập được là khoảng 2,3 triệu từ.	Viện Nghiên cứu Ngôn ngữ Quốc gia Nhật Bản, viết tắt là NINJAL). Đây là một cơ quan nghiên cứu trực thuộc Cơ quan Văn hóa Quốc gia Nhật Bản (独立行政法人国立文化財機構)
3	Khối ngữ liệu văn nói của người Nhật phiên bản 2018 (bản gồm cả phần lời thoại và ghi âm) (Tên tiếng Nhật: BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2018 年版	BTSJ (Basic Transcription System for Japanese) là một khối ngữ liệu lớn về hội thoại tự nhiên của người Nhật bản ngữ. Đây là nguồn dữ liệu hữu ích dành cho các nghiên cứu về ngữ dụng học, phân tích diễn ngôn, học máy, và giáo dục tiếng Nhật. Phiên bản năm 2018 của BTSJ Corpus là bản cập nhật mở rộng của phiên bản trước, bao gồm: Bổ sung thêm các đoạn hội thoại mới, cập nhật phiên bản phiên âm theo tiêu chuẩn năm 2015, bổ sung âm thanh cho các đoạn hội thoại trước đó chỉ có phiên âm.	Viện Nghiên cứu Quốc ngữ Quốc gia, Tổ chức Nghiên cứu Văn hóa Con người, Pháp nhân Tổ chức sử dụng chung giữa các đại học
4	Khối ngữ liệu khẩu ngữ tiếng Nhật (Tên tiếng Nhật: 日本語話し言葉コーパス	Đây là một khối ngữ liệu được xây dựng hoàn toàn là khẩu ngữ trong tiếng Nhật, dữ liệu được thu thập từ lời thoại, phát ngôn tiếng Nhật tự phát (đa số là độc thoại) và được chú giải (gán nhãn) rất chi tiết. Khối ngữ liệu này được công bố vào năm 2004 và đến nay vẫn duy trì chất lượng và quy mô ở mức hàng đầu thế giới. Tổng số từ được thu thập là khoảng 7,5 triệu từ. Khối ngữ liệu này hiện được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau như: Xử lý thông tin ngôn ngữ nói, Xử lý ngôn ngữ tự nhiên, Ngôn ngữ học tiếng Nhật, Ngôn ngữ học nói chung…	Viện Quốc gia về Ngôn ngữ học Nhật Bản, Viện Công nghệ Thông tin và Truyền thông (trước đây là Viện Nghiên cứu Truyền thông Tổng hợp) và Đại học Công nghệ Tokyo.
5	Khối ngữ liệu hội thoại của người Nhật bản ngữ do đại học Nagoya xây dựng (Tên tiếng Nhật: 名古屋大学会話コーパス）	Đây là một khối ngữ liệu hội thoại giữa người Nhật với nhau, được giáo sư Mieko Ōso thuộc Đại học Nagoya xây dựng bằng nguồn kinh phí nghiên cứu khoa học (Kakenhi) trong giai đoạn 2001 đến 2003. Ngữ liệu này gồm 129 cuộc hội thoại, với tổng thời lượng khoảng 100 giờ trò chuyện tự do. Tổng số từ lên đến khoảng 1,1 triệu từ. Có thể tra cứu bằng hệ thống 'Chūnagon' (中納言) qua trang web.	Đại học Nagoya Nhật Bản

6	Khối ngữ liệu "Taiyō" & Khối ngữ liệu tạp chí phụ nữ cận đại. (Tên tiếng Nhật: 太陽コーパス」と「近代女性雑誌コーパス」)	Khối ngữ liệu "Taiyō" và khối ngữ liệu tạp chí phụ nữ cận đại được công bố vào năm 2005, là khối ngữ liệu được xây dựng từ tạp chí tổng hợp *Taiyō* do nhà xuất bản Hakubunkan phát hành, một tạp chí từng được đọc rộng rãi trong thời kỳ Minh Trị và Đại Chính của Nhật. Khối ngữ liệu này bao gồm khoảng 14,5 triệu ký tự văn bản từ 5 năm: 1895, 1901, 1909, 1917 và 1925. Các năm này được lựa chọn bắt đầu từ năm tạp chí ra đời (1895), và sau đó cứ cách 8 năm một lần kể từ năm 1901 – năm đầu tiên của thế kỷ 20. Do chỉ dựa trên một loại tạp chí, nên khối ngữ liệu này không thể đảm bảo tính đại diện như BCCWJ. Tuy nhiên, vì *Taiyō* chứa nhiều văn bản đa dạng từ các bài tiểu luận đến tiểu thuyết, với sự tham gia của nhiều tác giả đương thời, nên đây là tư liệu rất thích hợp để nghiên cứu tiếng Nhật thời kỳ đó. Khối ngữ liệu Taiyō đã được xuất bản dưới dạng CD-ROM bởi Nhà xuất bản Sanseido, và có thể tìm kiếm toàn văn thông qua công cụ kèm theo mang tên "Himawari". Như đã được chỉ rõ bằng số lượng ký tự, kho này không gán thông tin cấp từ vựng, nên là một khối ngữ liệu dùng để tra cứu bằng chuỗi ký tự. Tuy nhiên, các văn bản đã được cấu trúc hóa bằng XML, với thông tin về bài viết và chỉnh sửa được gắn thẻ.	Viện Nghiên cứu Ngôn ngữ Quốc gia Nhật Bản (国立国語研究所/NINJAL)
7	Khối ngữ liệu lịch sử tiếng Nhật (Tên tiếng Nhật: 日本語歴史コーパス)	Đây là một khối ngữ liệu lưu trữ các tài liệu ngôn ngữ chính yếu trong lịch sử tiếng Nhật, từ *Manyōshū* (Vạn Diệp Tập) cho đến đầu thời kỳ Shōwa. Tính đến tháng 2 năm 2019, số lượng từ được lưu trữ là khoảng 15,6 triệu từ. Có thể tra cứu khối ngữ liệu này thông qua trang web *Chūnagon*.	Viện Ngôn ngữ và Văn hóa Nhật Bản
8	Khối ngữ liệu tạp chí Meiroku (Tên tiếng Nhật: 明六雑誌コーパス)	Khối ngữ liệu tạp chí Meiroku là khối ngữ liệu có quy mô nhỏ, được công bố vào năm 2012, là khối ngữ liệu được xây dựng từ toàn bộ các số của tạp chí *Meiroku Zasshi*, xuất bản vào đầu thời kỳ Minh Trị – trước *Taiyō* – với khoảng 180.000 từ. Trong khi "Khối ngữ liệu Taiyō" chỉ cho phép tìm kiếm chuỗi ký tự, thì "Khối ngữ liệu Meiroku" lại được gán thông tin từ đơn vị nhỏ tương thích với BCCWJ. Đây là khối ngữ liệu tiếng Nhật cổ đại đầu tiên được công bố rộng rãi có gán thông tin cấp từ. Khối ngữ liệu này được liên kết với các tệp hình ảnh bản gốc do Viện Quốc gia Nghiên cứu Ngôn ngữ Nhật Bản lưu trữ, cho phép tham khảo nguyên bản từ kết quả tìm kiếm. Giống như "Khối ngữ liệu tạp chí phụ nữ cận đại", kho này được công bố theo giấy phép Creative Commons.	Viện Nghiên cứu Ngôn ngữ Quốc gia Nhật Bản (国立国語研究所, National Institute for Japanese Language and Linguistics - NINJAL). Các trường đại học và tổ chức nghiên cứu Nhật Bản chuyên về ngôn ngữ học và lịch sử ngôn ngữ Nhật

9	Khối ngữ liệu tập hợp các bài viết của người nước ngoài học tiếng Nhật (Tên tiếng Nhật: 日本語学習者作文コーパス)	Khối ngữ liệu tập hợp bài viết của người học tiếng Nhật được phát triển trong khuôn khổ đề án nghiên cứu do Trung tâm Nghiên cứu Nhật Bản Quốc tế, Đại học Ngoại ngữ Tokyo chủ trì. Tên dự án là "Nghiên cứu giáo dục tiếng Nhật dựa trên ngôn ngữ mẹ đẻ và khu vực xuất thân của người học tiếng Nhật" (2010–2012), Các bài viết thu thập đã được xử lý và chú thích bằng thông tin chỉnh sửa (sửa lỗi ngôn ngữ) và gắn thẻ lỗi (error tags), từ đó hình thành một cơ sở dữ liệu có chú thích ngôn ngữ học. Khối ngữ liệu này không chỉ phục vụ cho mục đích nghiên cứu lỗi ngôn ngữ trong quá trình tiếp thu tiếng Nhật như một ngoại ngữ, mà còn hỗ trợ phát triển phương pháp giảng dạy tiếng Nhật phù hợp với đặc điểm ngôn ngữ mẹ đẻ và khu vực văn hóa của người học.	Đại học Ngoại ngữ Tokyo
10	Khối ngữ liệu KY Corpus (Tên tiếng Nhật: KY コーパス)	Khối ngữ liệu KY Corpus (KY コーパス) là một bộ sưu tập dữ liệu ngôn ngữ học được xây dựng từ các bản ghi OPI (Oral Proficiency Interview) của những người học tiếng Nhật, nhằm phục vụ nghiên cứu về sự phát triển ngôn ngữ và phân tích ngữ nghĩa. KY Corpus được xây dựng từ 90 bản ghi OPI của những người học tiếng Nhật, bao gồm các nhóm người nói tiếng Trung Quốc, Anh và Hàn Quốc. Mỗi nhóm có 30 người, được phân loại theo trình độ: sơ cấp (5 người), trung cấp (10 người), cao cấp (10 người) và siêu cấp (5 người). Khối ngữ liệu này được sử dụng để nghiên cứu về sự phát triển ngôn ngữ, quá trình thụ đắc ngôn ngữ của người học.	

Như vậy, có thể thấy rằng mỗi một khối ngữ liệu tiếng Nhật đều sở hữu những đặc điểm riêng biệt, từ đó tạo nên tiềm năng khai thác và ứng dụng khác nhau, tùy thuộc vào mục đích nghiên cứu hoặc lĩnh vực sử dụng. Việc nhận diện và khai thác các đặc điểm này một cách có hệ thống không chỉ giúp nâng cao hiệu quả sử dụng ngữ liệu, mà còn mở rộng phạm vi ứng dụng sang nhiều lĩnh vực liên ngành. Từ sự đa dạng này có thể thấy rằng không có một khối ngữ liệu "toàn năng", mỗi corpus nên được lựa chọn và khai thác dựa trên sự phù hợp với mục tiêu nghiên cứu cụ thể, từ đó tối ưu hóa giá trị ứng dụng của nó trong cả môi trường học thuật lẫn thực tiễn xã hội.

4. Ứng dụng khối ngữ liệu trong việc giảng dạy và nghiên cứu tiếng Nhật cho người Việt Nam

Trong tiếng Nhật, các khối ngữ liệu (コーパス / corpus) ngày càng phong phú và đa dạng, đóng vai trò không thể thiếu trong nhiều lĩnh vực như: ngôn ngữ học, giáo dục tiếng Nhật như ngôn ngữ thứ hai (JSL/JFL), xử lý ngôn ngữ tự nhiên (NLP), và nghiên cứu văn hóa – xã hội… Sự phát triển mạnh mẽ của các loại ngữ liệu này phản ánh nhu cầu thực tiễn ngày càng lớn đối với các nguồn dữ liệu ngôn ngữ có cấu trúc, mang tính đại diện và có thể truy cập cho nghiên cứu và ứng dụng công nghệ.

Hiện nay ở Việt Nam, giáo viên giảng dạy tiếng Nhật biết đến khối ngữ liệu và sử dụng

nó như công cụ nghiên cứu hầu như rất ít. Vì vậy, việc đưa ra khả năng khai thác và ứng dụng của các khối ngữ liệu điện tử của Nhật là rất hữu ích, giúp cho giáo viên cũng như nhà nghiên cứu có thể kịp thời nắm bắt và cập nhật thông tin một cách chính xác. Sau khi tìm hiểu nghiên cứu và đánh giá tiềm năng khai, tác giả thấy rằng mỗi một khối ngữ liệu tiếng Nhật đều sở hữu những đặc điểm riêng biệt, từ đó tạo nên tiềm năng khai thác và ứng dụng khác nhau, tùy thuộc vào mục đích nghiên cứu hoặc lĩnh vực sử dụng. Việc nhận diện và khai thác các đặc điểm này một cách có hệ thống không chỉ giúp nâng cao hiệu quả sử dụng ngữ liệu, mà còn mở rộng phạm vi ứng dụng sang nhiều lĩnh vực liên ngành. Trên cơ sở đánh giá về tiềm năng khai thác và ứng dụng của các khối ngữ liệu học trong tiếng Nhật, tác giả đề xuất một số định hướng nhằm nâng cao hiệu quả sử dụng các nguồn ngữ liệu ngôn ngữ này trong nghiên cứu ngôn ngữ nói chung và giảng dạy tiếng Nhật cho người Việt Nam nói riêng. Tiềm năng khai thác và ứng dụng của khối ngữ liệu được mô tả trong bảng dưới đây.

Bảng 2

Tiềm năng khai thác và ứng dụng của từng khối ngữ liệu trong tiếng Nhật

STT	Khối ngữ liệu	Tiềm năng khai thác và ứng dụng của khối ngữ liệu
1	Khối ngữ liệu ngôn ngữ viết cân bằng trong tiếng Nhật hiện đại (Tên tiếng Nhật: 現代日本語書き言葉均衡コーパス (BCCWJ)	** Nghiên cứu khảo sát về tần suất sử dụng, xuất hiện của từ vựng và biểu hiện của chúng để phục vụ cho công tác biên soạn từ điển Nhật Việt* Phân tích tần suất sử dụng và mức độ xuất hiện của từ ngữ trong các thể loại khác nhau như sách, báo, blog, v.v. để đưa ra danh mục từ phổ biến phục vụ cho việc biên soạn từ điển. Nghiên cứu của nhóm các học giả Yukio Tono, Makoto Yamazaki, Kikuo Maekawa công bố vào năm 2013 đã sử dụng Corpus CSJ và BCCWJ phân tích tần suất sử dụng và mức độ xuất hiện của 5000 từ phổ biến trong tiếng Nhật đương đại, cả văn nói và văn viết để biên soạn từ điển tiếng Nhật. ** Nghiên cứu phong cách ngôn ngữ và thể loại văn bản phục vụ cho nghiên cứu phong cách học và ngôn ngữ học chức năng* Phân tích sự khác biệt về phong cách ngôn ngữ văn học, hành chính, báo chí, blog,… xác định đặc trưng ngữ pháp và từ vựng riêng biệt cho từng thể loại. Nghiên cứu của Satoshi Nambu công bố vào năm 2025 đã so sánh giữa văn viết và văn nói, các thể loại trong BCCWJ, ảnh hưởng của các yếu tố ngoài ngôn ngữ (người nói, thể loại, lịch sử) lên việc chọn hình thức ngôn ngữ… ** Biên soạn giáo trình và giảng dạy tiếng Nhật cho người Việt Nam* Dựa vào tần suất và ngữ cảnh sử dụng từ, cụm từ trong văn bản thực tế để lựa chọn từ vựng cốt lõi cho giáo trình, thiết kế bài học theo ngữ cảnh sử dụng thực tế. Nghiên cứu của nhóm học giả Trường Đại học Ngoại ngữ Tokyo công bố vào năm 2008 đã sử dụng dữ liệu từ corpus phục vụ việc chọn 10.000 từ vựng cơ bản đọc hiểu để biên soạn giáo trình đọc hiểu cho người nước ngoài học tiếng Nhật. ** Nghiên cứu về so sánh đồng đại và lịch đại (ngôn ngữ học đồng đại & lịch đại) dành cho các điều tra khảo sát về ngôn ngữ học lịch sử, ngôn ngữ học xã hội.* Dùng khối ngữ liệu này để so sánh với các ngữ liệu cổ hơn như Khối ngữ liệu lịch sử tiếng Nhật/日本語歴史コーパス(CHJ) để phân tích sự thay đổi của từ vựng, ngữ pháp qua thời gian, đồng thời, có thể nghiên cứu ngôn ngữ hiện đại trong mối quan hệ với lịch sử tiếng Nhật. Nghiên cứu của Maekawa Kikuo và Ogiso Tomomi công bố năm 2020 đã sử dụng Corpus để khảo sát sự thay đổi về ngôn ngữ giữa ngữ liệu cổ (CHJ) và hiện đại (BCCWJ).

		** Tạo ngân hàng câu hỏi các dạng bài ôn thi JLPT sát thực tế ngôn ngữ của người bản ngữ* Sử dụng các ngữ liệu thực tế từ khối ngữ liệu BCCWJ làm nền để tạo câu hỏi, các dạng bài tập về ngữ pháp theo các cấp độ từ N5 đến N1. Nghiên cứu của Ishikawa Shinichiro công bố năm 2025 đã sử dụng Corpus để xử lý bằng hệ thống phân tích hình thái, thống kê tần suất xuất hiện của từ vựng để làm cơ sở cho bảng từ JLPT… ** Xây dựng tài liệu đọc/nghe dựa trên ngữ liệu thực tế từ BCCWJ* Biên soạn các dạng bài đọc và nghe dành cho cấp độ từ dễ đến khó phù hợp với trình độ của người học dựa trên ngữ liệu thực tế người bản ngữ sử dụng. Nghiên cứu của Matsushita công bố năm 2015 đã sử dụng Corpus để phân tích độ khó (tần suất, độ dài câu, số lượng từ Hán tự) của các đoạn văn trong corpus → phân cấp từ N5 đến N1 → tạo tài liệu đọc hiểu cho người học tiếng Nhật. Nghiên cứu của Kobayashi công bố năm 2018 đã sử dụng corpus để so sánh đặc trưng ngôn ngữ giữa ngôn ngữ viết (BCCWJ) và nói (CSJ), từ đó biên soạn tài liệu nghe thực tế phản ánh đặc điểm hội thoại tự nhiên (dạng "shadowing", "role-play").
2	Khối ngữ liệu văn nói của người học tiếng Nhật đến nhiều quốc gia có tiếng mẹ đẻ khác nhau (Tên tiếng Nhật: 多言語母語の日本語学習者横断コーパス)	** Nghiên cứu về lỗi sai trong cách dùng tiếng Nhật của người học (Error Analysis)* Dựa trên dữ liệu văn nói của người học để phân tích lỗi trong phát âm, ngữ pháp, cách dùng từ của người học theo từng ngôn ngữ mẹ đẻ của người học. Nghiên cứu của Sasano và Kato công bố năm 2020 đã sử dụng Corpus để thống kê và phát hiện lỗi trong bài viết người học tiếng Nhật… ** Nghiên cứu về tác động và ảnh hưởng của ngôn ngữ mẹ đẻ (L1 Transfer Studies) đến cách dùng tiếng Nhật của người Việt Nam* Tìm hiểu mức độ ảnh hưởng của ngôn ngữ mẹ đẻ (L1) đến cách dùng tiếng Nhật (L2), của người Việt Nam học tiếng Nhật đặc biệt là trong các phạm trù về trật tự từ (word order), Cách dùng thể lịch sự (敬語), thể bị động (受け身) và sai khiến (使役)… Nghiên cứu của Lê Thị Thanh công bố năm 2023 đã sử dụng Corpus để phân tích cách người Việt dùng sai 「～られる」 trong ngữ cảnh bị động, thường lược bỏ hoặc chuyển thành câu chủ động, do tiếng Việt không có biến thể bị động hình thái… ** Nghiên cứu so sánh về quá trình thụ đắc tiếng Nhật của người Việt Nam học tiếng Nhật với người học tiếng Nhật ở các nước khác* Trên cơ sở làm rõ quá trình thụ đắc tiếng Nhật của người Việt Nam, cũng như của người học tiếng Nhật ở các nước khác (Trung Quốc, Hàn Quốc, Thái Lan hay Anh…) để có thể nhận diện những đặc điểm nổi bật trong hành vi ngôn ngữ, chiến lược giao tiếp và khuynh hướng các lỗi sai điển hình trong cách dùng tiếng Nhật của người học. Nghiên cứu của Kobayashi và Ishikawa công bố năm 2025 đã sử dụng Corpus để so sánh ảnh hưởng của L1 trong các nhóm học viên (Việt, Hàn, Trung) trong việc thụ đắc ngữ pháp tiếng Nhật… ** Xây dựng chương trình học và giáo trình cho người Việt Nam học tiếng Nhật* Trên cơ sở phân tích kết quả sử dụng ngôn ngữ và xu hướng lỗi sai điển hình của người Việt Nam trong quá trình học tiếng Nhật, việc thiết kế giáo trình và tài liệu luyện tập chuyên biệt là cần thiết nhằm nâng cao hiệu quả tiếp nhận và vận dụng ngôn ngữ. Cụ thể, nội dung giảng dạy cần được biên soạn theo hướng phù hợp với đặc điểm ngôn ngữ mẹ đẻ của người học, đồng thời tập trung khắc phục các điểm yếu có tính hệ thống như lỗi ngữ pháp, lựa chọn từ vựng và phát âm. Nghiên cứu cụ thể của Nguyễn

		Thị Thu Minh công bố năm 2019 sử dụng Corpus để phân tích tần suất xuất hiện của từ vựng trong các thể loại khác nhau để thiết kế danh mục từ vựng cơ bản - trung cấp phù hợp cho giáo trình tiếng Nhật dành cho người Việt.
3	Khối ngữ liệu văn nói của người Nhật phiên bản 2018 (bản gồm cả phần lời thoại và ghi âm) (Tên tiếng Nhật: BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2018 年版	** Nghiên cứu về phân tích hội thoại của người bản ngữ* Nghiên cứu cách người Nhật thực hiện lời nói, ngắt quãng, phản hồi và xử lý sự bất đồng cũng như việc bắt đầu, duy trì, chuyển lượt và kết thúc trong hội thoại như thế nào… Nghiên cứu của Maekawa công bố năm 2020 đã sử dụng Corpus để gắn nhãn chức năng hội thoại trong CSJ (ví dụ: lời mở đầu, phản hồi, kết thúc), phục vụ nghiên cứu hội thoại và dạy tiếng Nhật giao tiếp. ** Nghiên cứu trong lĩnh vực ngữ dụng học (Pragmatics)* Sử dụng khối ngữ liệu để phân tích các chiến lược người Nhật sử dụng ngôn ngữ để đạt mục đích giao tiếp trong việc yêu cầu, từ chối, xin lỗi, cảm ơn… hoặc các nghiên cứu khảo sát về cách thức biểu đạt lịch sự, khiêm nhường (謙譲語) và tôn kính (尊敬語) của người Nhật. Nghiên cứu của Taniguchi công bố năm 2020 đã phân tích sự biến đổi của biểu thức xin lỗi theo lứa tuổi và giới tính… ** Nghiên cứu trong lĩnh vực ngôn ngữ ứng dụng* Xây dựng giáo trình, bài học dựa trên ngôn ngữ thực tế; biên soạn và thiết kế nội dung giảng dạy (các tình huống đối thoại thực tế) cho người Việt Nam học tiếng Nhật. Nghiên cứu của Nguyễn Thị Hồng và Phạm Thị Hương công bố năm 2021 đã dùng dữ liệu Corpus phân tích tần suất và ngữ cảnh sử dụng của các cấu trúc như「～ている」,「～そうだ」,「～ように」 trong ngữ liệu tiếng Nhật hiện đại, để xây dựng các bài học chú trọng ngữ dụng và tình huống giao tiếp thực tế, giúp người Việt học cách dùng biểu hiện tự nhiên. ** Nghiên cứu trong lĩnh vực Khoa học máy tính & Trí tuệ nhân tạo* Dữ liệu âm thanh chuẩn của người bản ngữ sẽ giúp huấn luyện mô hình AI cho nhận diện tiếng Nhật, ngoài ra, có thể giúp làm tăng độ chính xác cho trợ lý ảo, chatbot, dịch vụ thoại như: Siri, Google Assistant (tiếng Nhật). Dự án nghiên cứu của tổ chức NINJAL công bố năm 2022 đã dùng ngữ liệu phát âm người bản ngữ để huấn luyện mô hình AI giọng nói tiếng Nhật chuẩn, phục vụ phát triển dịch vụ thoại thông minh, ứng dụng chatbot giáo dục tiếng Nhật, nghiên cứu về cảm xúc và ngữ điệu trong tiếng Nhật. ** Nghiên cứu về các quy trình xử lý ngôn ngữ tự nhiên (NLP)* Phát triển các hệ thống dịch tự động, phân tích cảm xúc, phát hiện từ đồng âm/đa nghĩa trong tiếng Nhật. Nghiên cứu của Fujita công bố năm 2023 đã sử dụng Corpus để xây dựng hệ thống dịch tự động Nhật -Việt bằng AI, ứng dụng transformer models (MarianNMT, T5) được huấn luyện trên corpus thật… ** Nghiên cứu trong lĩnh vực ngôn ngữ học xã hội (Sociolinguistics)* Có thể tiến hành nghiên cứu và làm rõ sự khác biệt trong cách nói của các nhóm tuổi, giới tính, vùng miền (phương ngữ như Kansai-ben); tìm hiểu cách người trẻ nói khác người già, cách nói ở thành thị so với nông thôn; phân tích khảo sát về ngữ âm, ngữ điệu, trọng âm theo đặc điểm vùng miền của người bản ngữ. Dự án nghiên cứu của NINJAL công bố năm 2022 đã nghiên cứu biến thể ngôn ngữ theo vùng miền, độ tuổi, giới tính và môi trường xã hội để nhằm cung cấp dữ liệu cho các nghiên cứu sociophonetics và ngôn ngữ học biến thể.

4	Khối ngữ liệu khẩu ngữ tiếng Nhật (Tên tiếng Nhật: 日本語話し言葉コーパス)	** Nghiên cứu liên quan đến việc nhận dạng và xử lý tiếng nói (Speech Recognition/ASR)* Đây là nguồn dữ liệu nền tảng để huấn luyện các hệ thống nhận diện tiếng Nhật tự động. Ví dụ, thông qua các bài diễn thuyết dài và liên tục sẽ giúp mô phỏng tốt các tình huống hội thoại thực tế. Dữ liệu bao gồm: âm thanh, bản chép lời (transcripts), thông tin ngữ âm và ngữ điệu nên rất lý tưởng cho học máy, có thể được sử dụng trong các hệ thống chuyển giọng nói thành văn bản (speech-to-text) như: Google Voice hoặc Siri tiếng Nhật. Nghiên cứu của Yamagishi và Kobayashi công bố năm 2023 đã xây dựng corpus giọng nói cảm xúc tiếng Nhật, ứng dụng vào ASR nhận biết cảm xúc và TTS (Text-to-Speech). ** Nghiên cứu liên quan đến phân tích cú pháp và ngữ pháp trong lời nói tự nhiên* Dữ liệu sẽ giúp phân tích các đặc điểm cú pháp khi con người nói chuyện, phục vụ nghiên cứu ngôn ngữ học thực chứng (empirical linguistics), tạo nền tảng cho lý thuyết về ngôn ngữ nói. Nghiên cứu của Koiso và Maekawa công bố năm 2025 đã sử dụng corpus để phân tích mối quan hệ giữa ngữ pháp (syntax) và ngữ điệu (prosody) trong phát ngôn tự nhiên, dữ liệu được gắn nhãn cú pháp, ranh giới phát ngôn và ngữ điệu… ** Nghiên cứu liên quan đến phân tích hội thoại và ngữ dụng học* Khối ngữ liệu cung cấp dữ liệu hữu ích cho việc nghiên cứu về các chiến lược giữ lượt nói, chuyển lượt; cách sử dụng đệm từ (フィラー) như「えーと」「あのー」hoặc các dạng thức biểu hiện như 「～じゃん」「～っけ」「～さ」 vốn khó tìm thấy trong văn viết. Nghiên cứu của Koiso công bố năm 2021 đã mô hình hóa cơ chế giữ lượt nói (floor-holding) bằng filler「あの」「えっと」để làm rõ chiến lược phản ứng linh hoạt của người Nhật với đối phương bằng aizuchi thể hiện sự chú ý và đồng thuận… ** Nghiên cứu về xây dựng biên soạn giáo trình nghe nói tiếng Nhật* Có thể dựa vào dữ liệu để soạn bài nghe cho kỳ thi JLPT N2/N1, hay luyện nói trong các tình huống thực tế như: hội họp, phát biểu, phỏng vấn. Dự án nghiên cứu của NINJAL công bố năm 2020 đã sử dụng corpus để xây dựng ngân hàng dữ liệu âm thanh gồm các đoạn hội thoại ngắn, có kịch bản và transcript, phân cấp theo trình độ để làm năng lực JLPT…
5	Khối ngữ liệu hội thoại của Đại học Nagoya (Tên tiếng Nhật: 名古屋大学会話コーパス）	** Nghiên cứu về phân tích hội thoại (Conversation Analysis)* Có thể tiến hành các điều tra nghiên cứu về cấu trúc và tổ chức cuộc hội thoại trong ngôn ngữ nói tự nhiên; hoặc nghiên cứu về các chiến lược chuyển lượt nói (turn-taking), phản hồi ngầm (backchannel), và cách xử lý ngắt lời, sửa lỗi trong giao tiếp. Hơn nữa, có thể khảo sát về phân tích các biểu hiện ngôn ngữ trong tương tác xã hội như biểu hiện lịch sự, sự đồng thuận hay bất đồng. Nghiên cứu của Kawamori công bố năm 2019, sử dụng Corpus để phân tích phản hồi ngầm (aizuchi) trong các cuộc trò chuyện tự nhiên, đo tần suất, loại hình phản hồi (“うん”, “はい”, “へぇ”, “そうなんだ”) và mối liên hệ với giới tính, độ tuổi. ** Nghiên cứu về ngữ âm và ngữ điệu của tiếng Nhật* Khảo sát điều tra để làm rõ đặc điểm phát âm, ngữ điệu, cao độ (pitch), ngắt nghỉ trong hội thoại tự nhiên của người Nhật; hoặc so sánh các đặc trưng ngữ âm giữa các nhóm người nói, như khác biệt vùng miền hay giới tính. Nghiên cứu của Nakamura và cộng sự công bố năm 2019 sử dụng corpus *Japanese Emotional Speech Database (JTES)* để huấn luyện mô hình nhận diện cảm xúc trong phát ngôn tiếng Nhật.

		** Ứng dụng trong công nghệ xử lý ngôn ngữ tự nhiên (NLP)* Có thể ứng dụng trong lĩnh vực huấn luyện và kiểm thử các hệ thống nhận dạng tiếng nói tự động (ASR) bằng dữ liệu hội thoại thực tế; phát triển các chatbot và trợ lý ảo tiếng Nhật có khả năng xử lý hội thoại tự nhiên; nghiên cứu và phát triển các công cụ phân tích cảm xúc và tương tác xã hội trong hội thoại. ** Nghiên cứu liên quan đến các vấn đề tương tác xã hội và giao tiếp phi ngôn ngữ* Khảo sát các biểu hiện phi ngôn ngữ (như ngôn ngữ cơ thể, ánh mắt, cử chỉ) kết hợp với lời nói trong hội thoại; phân tích sự phối hợp giữa lời nói và tín hiệu phi ngôn ngữ trong việc xây dựng mối quan hệ xã hội.
6	Kho ngữ liệu "Taiyō" & Kho ngữ liệu tạp chí phụ nữ cận đại. (Tên tiếng Nhật: 太陽コーパス」と「近代女性雑誌コーパス」)	** Nghiên cứu ngôn ngữ học lịch sử (Historical Linguistics)* Nghiên cứu ngôn ngữ học lịch sử để làm rõ sự thay đổi trong từ vựng, ngữ pháp, phong cách văn viết từ cuối thế kỷ 19 đến đầu thế kỷ 20. Ví dụ: nghiên cứu về chủ đề so sánh cách dùng kính ngữ trong tạp chí phụ nữ với văn học đại chúng trong "Taiyō"… ** Nghiên cứu về xã hội và văn hóa Nhật Bản* Kho ngữ liệu Tạp chí phụ nữ cận đại chứa các bài viết về vai trò giới, đạo đức, hôn nhân, và gia đình, phản ánh quan niệm truyền thống với hiện đại hóa. Kho ngữ liệu "Taiyō" phản ánh các vấn đề về chính sách quốc gia, quan điểm về chế độ, hiện đại hóa và quan hệ quốc tế. Do vậy, có thể tiến hành các nghiên cứu liên quan đến các vấn đề ý thức hệ của người Nhật trong xã hội xưa và nay; các khảo sát điều tra để làm rõ tư tưởng cũng như cách nhìn nhận của xã hội về giới tính, gia đình, giáo dục, mô hình lý tưởng phụ nữ. ** Ứng dụng trong công nghệ ngôn ngữ và AI* Là nguồn dữ liệu huấn luyện cho các mô hình ngôn ngữ xử lý tiếng Nhật cổ - cận đại, dùng trong nghiên cứu về OCR (nhận diện ký tự), phục dựng tài liệu cổ, dịch máy học thuật...
7	Kho ngữ liệu tạp chí Meiroku (Tên tiếng Nhật: 明六雑誌コーパス)	** Nghiên cứu về tư tưởng và triết học cận đại Nhật Bản* Kho ngữ liệu tạp chí Meiroku phản ánh các tư tưởng khai sáng của phương Tây như: tự do, bình đẳng, dân chủ, khoa học, giáo dục, và pháp trị, nên kho ngữ liệu này cho phép truy xuất hệ thống hóa các chủ đề triết học – chính trị và tư tưởng cá nhân của từng trí thức. ** Nghiên cứu về phân tích ngôn ngữ học lịch sử* Kho ngữ liệu cung cấp dữ liệu ngôn ngữ chuẩn của văn viết đầu thời Meiji, hữu ích trong việc nghiên cứu quá trình chuyển biến từ văn ngôn (漢文訓読) sang văn bạch thoại (口語文); hoặc các đề tài nghiên cứu có liên quan đến lịch sử ra đời và phát triển của từ ngoại lai trong tiếng Nhật… ** Nghiên cứu về quá trình tiếp biến văn hóa – khái niệm hiện đại* Phân tích cách các khái niệm phương Tây được dịch, định nghĩa và tranh luận trong xã hội Nhật đầu thời Minh Trị, dùng để truy tìm sự hình thành và phát triển của từ mượn phương Tây (gairaigo 外来語) và từ Hán mới tạo (wasei kango 和製漢語). ** Ứng dụng trong giáo dục, lịch sử tư tưởng và xã hội học* Hữu ích trong các chương trình giảng dạy về lịch sử tư tưởng Nhật Bản, Đông Á hiện đại, và các môn học về chuyển đổi xã hội; giúp sinh viên và nhà nghiên cứu tiếp cận trực tiếp nguồn văn bản gốc để hiểu sâu sắc bối cảnh lịch sử – tư tưởng.

Có thể thấy rằng corpus đóng vai trò quan trọng trong nhiều lĩnh vực liên quan đến tiếng Nhật. Trong nghiên cứu ngôn ngữ học, corpus giúp phân tích cách sử dụng từ vựng, ngữ pháp

và kính ngữ trong thực tế, từ đó đưa ra cái nhìn chính xác hơn về tiếng Nhật hiện đại… Trong giảng dạy tiếng Nhật, nó hỗ trợ xây dựng giáo trình, từ điển và tài liệu học tập dựa trên ngôn ngữ được sử dụng thực tế, giúp người học tiếp cận tiếng Nhật một cách tự nhiên và hiệu quả hơn… Ngoài ra, corpus còn được ứng dụng rộng rãi trong lĩnh vực công nghệ như: xử lý ngôn ngữ tự nhiên, dịch máy, nhận diện giọng nói và phát triển các hệ thống trí tuệ nhân tạo liên quan đến tiếng Nhật… Nhờ những ứng dụng đa dạng đó, corpus trở thành công cụ không thể thiếu trong việc nghiên cứu và ứng dụng tiếng Nhật hiện đại.

5. Thảo luận và đề xuất

Như vậy, chúng ta có thể khẳng định rằng việc khai thác và ứng dụng ngữ liệu điện tử trong tiếng Nhật có ý nghĩa rất thiết thực và hữu ích đối với việc giảng dạy và nghiên cứu tiếng Nhật cho người Viêt Nam. Cụ thể, nó có thể ứng dụng vào: Biên soạn tài liệu giảng dạy dựa trên dữ liệu ngôn ngữ thực tế; tìm hiểu cách sử dụng từ và cấu trúc câu trong ngữ cảnh thực tế; giảng dạy ngữ pháp và từ vựng theo các ngữ cảnh xuất hiện trong thực tiễn; phân tích tần suất từ vựng và mẫu câu; phân tích lỗi sai của người học; nghiên cứu quá trình thụ đắc ngôn ngữ thứ hai (Second Language Acquisition - SLA); nghiên cứu biến thể phương ngữ và ngôn ngữ nói; so sánh sự khác biệt giữa tiếng Nhật hiện đại và lịch sử; nghiên cứu biến đổi ngôn ngữ theo thời gian và không gian; so sánh đối chiếu ngôn ngữ học; nghiên cứu về phân tích ngôn ngữ học lịch sử; nghiên cứu về tư tưởng và triết học cận đại Nhật Bản; nghiên cứu về xã hội văn hóa, văn học Nhật Bản; nghiên cứu về so sánh phương ngữ và ảnh hưởng vùng miền; ứng dụng vào công nghệ ngôn ngữ và trí tuệ nhân tạo… Tuy nhiên, hiện nay, việc ứng dụng này chưa thật sự phổ biến và chưa đạt được mục đích mong muốn bởi lẽ những kho ngữ liệu tiếng Nhật và ngữ liệu đa ngôn ngữ có chứa tiếng Việt chưa nhiều và chưa được phổ biến rộng rãi. Hầu hết các kho ngữ liệu có gắn nhãn đều không được cung cấp miễn phí, người sử dụng muốn sử dụng phải mua với giá khá cao. Ngoài ra, hầu hết các kho ngữ liệu có chứa tiếng Nhật hiện có đều là kho ngữ liệu phục vụ cho nhiều mục đích khác nhau chứ chưa có những kho ngữ liệu chuyên biệt chỉ phục vụ cho việc giảng dạy tiếng Nhật cho người nước ngoài. Tiếp đến, việc khai thác, sử dụng hiện quả các kho ngữ liệu cũng cần đòi hỏi người sử dụng phải có những kiến thức cơ bản về công nghệ thông tin do chúng thường được đọc bởi một phần mềm hay một công cụ tìm kiếm nhất định. Chính vì vậy, tác giả xin nêu một số đề xuất nhằm mục đích nâng cao tính hiệu quả của việc khai thác và ứng dụng được những kho ngữ liệu điện tử trong việc giảng dạy tiếng Nhật cho người Việt Nam ở phần dưới đây:

- Thứ nhất, cần xây dựng thêm những khối ngữ liệu chuyên biệt phục vụ cho những mục đích giảng dạy tiếng Nhật cụ thể. Chẳng hạn như: những khối ngữ liệu phục vụ cho việc biên soạn giáo trình tiếng Nhật cho người Việt Nam (khối ngữ liệu các bài text, các bài hội thoại, các mẫu câu theo từng trình độ khác nhau), những khối ngữ liệu Nhật chuyên ngành (chuyên ngành du lịch, chuyên ngành thương mại, chuyên ngành hành chính, văn phòng…), những khối ngữ liệu ngân hàng đề thi đánh giá năng lực tiếng Nhật theo từng cấp độ, những khối ngữ liệu văn bản nói (hội thoại, thuyết trình, bài giảng, bản tin...) phục vụ cho việc giảng dạy tiếng Nhật tại Việt Nam.

- Thứ hai, đội ngũ giảng viên phải không ngừng học tập để nâng cao trình độ sử dụng công nghệ thông tin, có như vậy mới khai thác hiệu quả các khối ngữ liệu điện tử. Hiện nay, ngoài những kho ngữ liệu tiếng Nhật do người Nhật xây dựng còn có những kho ngữ liệu tiếng Nhật hoặc ngữ liệu song ngữ chứa tiếng Nhật do các tổ chức nước ngoài xây dựng. Muốn sử dụng được chúng, cần thiết phải sử dụng được một số phần mềm và công cụ tìm kiếm ngữ liệu trực tuyến như: AntConc, KH Coder, hoặc LancsBox cũng rất hữu ích trong việc khai thác và

xử lý ngữ liệu phục vụ giảng dạy và nghiên cứu.

- Thứ ba, cần triển khai những đề tài hoặc dự án nghiên cứu về việc khai thác những khối ngữ liệu điện tử để xây dựng các khóa học tiếng Nhật online. Khối ngữ liệu không chỉ giúp ích trong việc xây dựng giáo trình bản giấy mà còn rất hữu ích trong việc xây dựng giáo trình online. Nhờ những khối ngữ liệu có sẵn, việc đưa nội dung bài giảng vào các chương trình học online sẽ tiết kiệm được rất nhiều thời gian và công sức của người dạy.

6. Kết luận

Trong bối cảnh cách mạng công nghiệp 4.0, sự phát triển mạnh mẽ của khoa học máy tính và trí tuệ nhân tạo đã tạo ra những bước chuyển mình lớn trong lĩnh vực ngôn ngữ học và giảng dạy ngoại ngữ. Một trong những thành tựu nổi bật là sự ra đời và phổ biến của các khối ngữ liệu điện tử - nguồn dữ liệu ngôn ngữ lớn được thu thập, xử lý và chú giải có hệ thống từ ngôn ngữ tự nhiên. Việc khai thác hiệu quả các corpus này đã chứng minh được giá trị to lớn trong giảng dạy và nghiên cứu tiếng Nhật, đặc biệt đối với người học là người Việt Nam. Tại Nhật Bản, nhiều công trình nghiên cứu đã ứng dụng corpus vào giảng dạy tiếng Nhật. Dữ liệu từ các kho này đã được sử dụng trong nhiều dự án biên soạn giáo trình, từ điển… Đối với người Việt Nam học tiếng Nhật, việc ứng dụng corpus mang lại nhiều lợi ích thiết thực. Có rất nhiều các nghiên cứu đã bước đầu thử nghiệm sử dụng dữ liệu từ BCCWJ và Leeds Japanese Corpus… để biên soạn bài đọc và bài luyện nghe, ngữ pháp… phù hợp với trình độ người học Việt Nam. Kết quả cho thấy việc sử dụng ví dụ, mẫu câu và đoạn hội thoại được trích xuất từ corpus giúp sinh viên tiếp cận ngôn ngữ tự nhiên, tăng khả năng hiểu ngữ cảnh và giảm lỗi ngữ dụng. Nhờ khả năng phản ánh trung thực ngôn ngữ trong đời sống, corpus không chỉ hỗ trợ nâng cao chất lượng giảng dạy mà còn mở rộng năng lực nghiên cứu cho giảng viên và học viên. Trong bối cảnh giáo dục ngôn ngữ hiện đại ngày càng hướng đến tính thực tiễn, dữ liệu hóa và cá nhân hóa, việc khai thác và ứng dụng hệ thống khối ngữ liệu điện tử tiếng Nhật một cách khoa học, có định hướng cho người học Việt Nam là xu thế tất yếu trong tương lai.

Tài liệu tham khảo

Baker, P. (2006). *Using corpora in discourse analysis*. Continuum.

Dao, H. T. (2007*).* Corpus linguistics (Part 1). *Journal of Language and Life, 7*(141).

Dinh, D. (2018). *Monograph on corpus linguistics*. Vietnam National University – Ho Chi Minh City Publishing House.

Dinh, D., & Ho, H. T. (2011). Computational linguistics and the development of dictionaries. *Journal of Lexicography & Encyclopedia, 4*(12)/7.

Dinh, D., & Ho, X. V. (2016). Applications of corpora in teaching Vietnamese to foreigners. In *Proceedings of the International Conference on Teaching and Researching Vietnamese Studies and the Vietnamese Language* (pp. 172-180).

Dinh, D., & Ly, N. M. (2015). The application of English–Vietnamese parallel corpora in language teaching. In *Proceedings of the Interdisciplinary Conference on Applied Linguistics & Language Teaching* (pp. 559-567).

Ide, N., & Suderman, K. (2007). The open corpus initiative for Asian languages. *Language Resources and Evaluation, 41*(3–4), 397-414.

Ikehara, S., Isahara, H., & Kurohashi, S. (2004). Balanced corpus of contemporary written Japanese. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)* (pp. 1271-1274).

Kudo, T., & Matsumoto, Y. (2002). Construction of Japanese corpus and its application to language processing. In *Proceedings of the 19th International Conference on Computational Linguistics (Vol. 1)* (pp. 1-7).

Kurohashi, S., & Nagao, M. (1997). Building a Japanese parsed corpus and its application to NLP. In *Proceedings*

of the 5th Workshop on Very Large Corpora (pp. 12-18).

Lam, T. T. (2022). Exploiting and applying electronic corpora in teaching Vietnamese to foreigners. In *Proceedings of the National Conference, Hue University* (pp. 70-78).

Le, T. N. (2020). Corpus linguistics – Concepts, approaches, methods, and applications in research and teaching German as a foreign language. *VNU Journal of Foreign Studies, 36*(5), 75-90. https://doi.org/10.25073/2525-2445/vnufs.4609

Maekawa, K., Nakatani, T., & Ogura, Y. (2014). Design and construction of the Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation, 48*(2), 345-371.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Wades, S. E., & Moje, E. B. (2000). *The role of text in classroom learning.* In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research (Vol. 3)* (pp. 609-627).