XÂY DỰNG KHO NGỮ LIỆU PHỤC VỤ DẠY TIẾNG VIỆT CHO NGƯỜI NƯỚC NGOÀI

VÕ DIỆP NHƯ[1]
BÙI THANH TƯỜNG THUỴ[2]
PHAN THANH BẢO TRÂN[3]
ĐINH ĐIỀN[4]

Abstract: In teaching Vietnamese to foreigners, there are various methods used to increase the efficiency and help learners access Vietnamese easily. The application of corpora in teaching is one of the useful and practical methods, as it increases the acquisition and usage of non-native languages. However, in Vietnam, this method still faces numerous limitations due to the lack of suitable corpora that meet the needs of teaching and learning Vietnamese. In this paper, we discuss the procedure of building a Vietnamese corpus, sampling, collecting, and processing materials. This corpus is used to construct vocabulary sets appropriate for each grade level and extract examples of vocabulary usage for each real-world context.

Key words: *applied linguistics, corpus, corpus linguistics*.

1. Giới thiệu

Hiện nay, tiếng Việt đã trở thành một trong những ngôn ngữ được sử dụng phổ biến và là một trong 20 ngôn ngữ được sử dụng nhiều nhất trên toàn thế giới [22]. Số lượng người nước ngoài đến Việt Nam làm việc và sinh sống càng nhiều, họ càng quan tâm và có nhu cầu học tập, nghiên cứu tiếng Việt. Bên cạnh các nghiên cứu cũng như phương pháp tiếp cận mới trong vấn đề giảng dạy tiếng Việt cho người nước ngoài như: sử dụng phương pháp giao tiếp [1, tr. 68]; giảng dạy tiếng Việt qua thành ngữ, tục ngữ [3, tr. 81]; giảng dạy hành động ngôn từ tiếng Việt theo định hướng chức năng, kết hợp các vấn đề giao tiếp liên văn hoá [10, tr. 67-71];… tác giả Đinh Điền cũng đã trình bày một hướng tiếp cận mới trong lĩnh vực này đó chính là ứng dụng ngôn ngữ học tính toán trong việc giảng dạy tiếng Việt cho người nước ngoài [7, tr. 19], đặc biệt phải kể đến vai trò của việc ứng dụng ngôn ngữ học tính toán để rút trích từ các kết quả từ kho ngữ liệu phục vụ cho việc giảng dạy.

Trong phạm vi bài viết này, chúng tôi trình bày việc xây dựng kho ngữ liệu đơn ngữ tiếng Việt phiên bản đầu tiên (Vietnmese Corpus 1.0 - VnC 1.0) bao gồm văn bản nói và văn bản viết, được thu thập chủ yếu từ năm 2000 đến năm 2020. Kho ngữ liệu liệu này có kích thước lớn (100 triệu từ) đảm bảo được tính cân bằng và tính đại diện. Các kết quả rút trích từ kho ngữ liệu này mang nhiều giá trị trong việc ứng dụng dạy tiếng Việt cho người nước ngoài. Phần hai của bài viết trình bày các

[1, 4] Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh.
[2] Trường Đại học VinUni.
[3] Trường Đại học Khoa học Xã hội và Nhân văn, Đại học Quốc gia Thành phố Hồ Chí Minh.

công trình liên quan đến việc sử dụng kho ngữ liệu trong giảng dạy ngoại ngữ trên thế giới cũng như trong nước. Phần ba đề cập đến việc xây dựng kho ngữ liệu đơn ngữ tiếng Việt phục vụ dạy tiếng Việt cho người nước ngoài. Phần bốn trình bày các ứng dụng cụ thể của các kết quả rút trích từ kho ngữ liệu trong giảng dạy tiếng Việt và phần năm là phần kết luận.

2. Các công trình liên quan

Trên thế giới, việc giảng dạy ngôn ngữ nói chung và giảng dạy ngoại ngữ nói riêng với sự hỗ trợ của các kho ngữ liệu đã có bề dày lịch sử. Dự án COBUILD tại Đại học Birmingham vào năm 1987 đã xây dựng kho ngữ liệu được thiết kết đặc biệt với mục đích sư phạm, gồm từ điển với các mục từ được chọn và sắp xếp theo tần suất cùng các ví dụ sử dụng ngôn ngữ trong thực tế được rút trích từ kho ngữ liệu. Nhờ vào sự phát triển của công nghệ thông tin, việc ứng dụng danh sách tần số từ vào giảng dạy ngoại ngữ vẫn được áp dụng rộng rãi cho đến ngày nay, tạo tiền đề cho sự phát triển của những kho ngữ liệu điện tử có kích thước vô cùng lớn (lên đến hàng trăm triệu từ hoặc hàng tỉ từ). Điển hình là kho ngữ liệu Quốc gia Anh - British National Corpus (BNC) [14], kho ngữ liệu Corpus of Contemporary American English (COCA) [15, tr. 453]. Các kho ngữ liệu của các ngôn ngữ khác cũng dần ra đời như kho ngữ liệu tiếng Tây Ban Nha, Trung Quốc, Séc, Pháp, Nhật Bản, Hàn Quốc,… [25].

Năm 2001, nghiên cứu của Elke đã chỉ ra việc ứng dụng kho ngữ liệu song ngữ Đức - Anh INTERSECT trong giảng dạy tiếng Đức, mang lại hiệu quả cho những sinh viên mới bắt đầu học tiếng Đức, sử dụng chuỗi đồng hiện từ kho ngữ liệu mang lại nhiều lợi ích trong giảng dạy ngữ pháp cũng như tăng sự thú vị trong các bài tập về từ vựng. Sách ngữ pháp [21], tài liệu bổ trợ [19, tr. 29] đã được biên soạn dựa trên dữ liệu từ kho ngữ liệu cung cấp các thông tin đặc biệt phải kể đến là các ví dụ đồng hiện, hỗ trợ người học trong việc sử dụng để khám phá ngôn ngữ hoặc suy luận (kiểm tra quy tắc sử dụng ngôn ngữ, phân loại cách sử dụng ngôn ngữ). Các kho ngữ liệu cũng đã được sử dụng trong việc xây dựng các tài liệu giảng dạy như Touchstone 4 [17]. Chujio, Utiyama và Miura đã công bố kết quả thực nghiệm trong nghiên cứu của mình về việc ứng dụng kho ngữ liệu song ngữ Nhật - Anh trong môi trường giảng dạy tiếng Anh sơ cấp cho sinh viên; có đến 80% sinh viên phản hồi rằng họ mong muốn sử dụng công cụ khai thác chuỗi đồng hiện trong kho ngữ liệu một cách thường xuyên để hỗ trợ quá trình tự học tiếng Anh [13, tr.12]. Nghiên cứu của Cobb đã đưa ra vai trò của việc sử dụng kho ngữ liệu hỗ trợ giảng dạy ngôn ngữ theo ngữ cảnh [24, tr. 38]. Năm 2010, các công trình nghiên cứu của Reppen và Bennett đã chỉ ra sách ngữ pháp như LGSWE có sử dụng thông tin từ kho ngữ liệu giúp người học dễ tiếp thu các vấn đề về tần suất, hình thái, kết ngôn,… [12, tr. 27], [20, tr. 21]. Đến năm 2011, trong công trình của mình, Romer đã khẳng định lợi ích của việc sử dụng kho ngữ liệu trong biên soạn giáo trình cũng như giảng dạy ngôn ngữ thứ hai [27, tr. 205]. Nghiên cứu của Donesch-Jezo đã cung cấp các ví dụ về các hoạt động học tập cho thấy rằng cách giảng dạy từ vựng đối với ngôn ngữ thứ hai được thực hiện một cách hiệu quả và hấp dẫn nhờ vào việc sử dụng kho ngữ liệu và các chuỗi đồng hiện làm tư liệu giảng dạy [11, tr. 23]. Son đã nghiên cứu việc sử dụng kho ngữ liệu, tập trung vào khai thác chuỗi đồng hiện trong việc dạy và học ngoại ngữ (tiếng Anh) trong các lớp học tại Việt Nam, kết quả những học sinh có sự hỗ trợ của các chuỗi đồng hiện từ kho ngữ liệu có thể dễ dàng tiếp thu trong việc khám phá cấu trúc từ và các

quy tắc ngữ pháp, giúp các học sinh này trả lời đúng nhiều câu hỏi hơn so với những học sinh còn lại [28]. Năm 2020, Suhardijanto và Putra đã nghiên cứu hiệu quả của việc sử dụng kho ngữ liệu Korpus BIPA gồm hơn 2 triệu từ trong việc giảng dạy tiếng Indonesia như một ngoại ngữ. Kết quả cho thấy, người học tăng khả năng phân biệt khi sử dụng những đơn vị từ trong tiếng Indonesia mang nghĩa "thoải mái" như phân biệt cách sử dụng từ *baik* và *bagus* đồng thời năng lực của người học cũng được nâng cao [26, tr. 6].

Tại Việt Nam, có một số nghiên cứu tính ứng dụng của kho ngữ liệu trong giảng dạy tiếng Việt cho người nước ngoài như nghiên cứu của tác giả Đào Hồng Thu đã trình bày vai trò và một số giải pháp của việc sử dụng khối liệu ngôn ngữ (kho ngữ liệu) trong giảng dạy tiếng Việt [4, tr. 64]. Tác giả Phạm Hiển, trong nghiên cứu của mình năm 2012, đã giới thiệu và đề xuất việc sử dụng kho ngữ liệu trong quá trình dạy, học tiếng Việt. Trong đó, tác giả nhấn mạnh đến các phương pháp sử dụng kho ngữ liệu trong giảng dạy tiếng Việt như: lập danh sách tần số từ, sử dụng dòng chỉ mục ngữ cảnh (concordance), sử dụng bảng đồng hiện từ vựng (collocation), lập khuôn cấu trúc câu. Bên cạnh đó, tác giả cũng đưa ra một số ứng dụng thực tế của kho ngữ liệu trong việc biên soạn bài thực hành từ vựng, ngữ pháp [9, tr. 22-26]. Năm 2015, trong nghiên cứu về ứng dụng kho ngữ liệu song ngữ Anh - Việt trong giảng dạy ngôn ngữ, tác giả Đinh Điền và Lý Ngọc Minh đã khảo sát cách dùng từ qua chuỗi đồng hiện (concordance), khảo sát kết ngôn (collocation) và khai thác chuyển ngữ Anh - Việt cho một số cụm từ cố định từ kho ngữ liệu song ngữ Anh - Việt EVC phục vụ giảng dạy tiếng Việt cho người nước ngoài [6]. Trong công trình năm 2016, tác giả Đinh Điền và Hồ Xuân Vinh nghiên cứu việc ứng dụng kho ngữ liệu đơn ngữ trong việc dạy tiếng Việt cho người nước ngoài, các kết quả thống kê hai kho ngữ liệu VCor và VTB được khai thác phục vụ cho việc giảng dạy [5]. Tuy nhiên hai kho ngữ liệu VCor và VTB đến nay đang dừng lại ở việc thu thập ngữ liệu từ các trang báo điện tử. Vì thế, chúng tôi tiến hành xây dựng kho ngữ liệu tiếng Việt cân bằng, mang tính đại diện cho tiếng Việt. Kho ngữ liệu này sẽ gồm văn bản nói và văn bản viết, bên cạnh đó thu thập ngữ liệu theo nhiều lĩnh vực, các kết quả rút trích từ kho ngữ liệu chúng tôi đã xây dựng sẽ đóng góp thêm nhiều thông tin ngôn ngữ, rút trích được nhiều ví dụ sử dụng ngôn ngữ trong nhiều lĩnh vực.

3. Xây dựng kho ngữ liệu tiếng Việt (VnC 1.0)

Trong phần này, chúng tôi sẽ trình bày vấn đề xây dựng kho ngữ liệu tiếng Việt phục vụ dạy tiếng Việt cho người nước ngoài, cụ thể là việc xây dựng kho ngữ liệu VnC 1.0.

3.1. Lập kế hoạch

Việc lập kế hoạch trước khi xây dựng kho ngữ liệu giúp những người xây dựng có cái nhìn tổng quan, xác định mục tiêu về thiết kế ngôn ngữ cũng như việc quản lí chi phí trong suốt quá trình xây dựng kho ngữ liệu [23, tr. 3].

Các mục tiêu về thiết kế ngôn ngữ được chúng tôi đặt ra như sau:

- Loại kho ngữ liệu sẽ xây dựng: kho ngữ liệu đơn ngữ, gồm ngữ liệu chuyên ngành và phổ thông theo tỉ lệ nhất định, thu thập chủ yếu các văn bản đương đại.
- Kích thước của mỗi mẫu văn bản: kích thước mẫu văn bản phụ thuộc vào loại văn bản mà chúng tôi thu thập, ví dụ văn bản báo điện tử sẽ chứa vài trăm từ nhưng văn bản truyện ngắn sẽ chứa

vài chục nghìn từ. Tuy nhiên, không thể lấy mẫu văn bản quá lớn, điều này sẽ làm ảnh hưởng đến kết quả khai thác, đối với những văn bản lớn như kịch bản kịch sân khấu, chúng tôi chỉ rút trích một phần.

- Kho ngữ liệu được xây dựng và thu thập văn bản theo các biến thể ngôn ngữ (đồng đại).

- Dạng thức văn bản: văn bản viết và văn bản nói.

Các mục tiêu về quản lí chi phí mà chúng tôi cần xem xét đến khi lập kết hoạch như:

- Số người thu thập, xử lí, đánh giá ngữ liệu
- Các chi phí phát sinh trong quá trình xây dựng: chi phí về phần cứng, chi phí mua hoặc xây dựng phần mềm xử lí ngữ liệu, chi phí thuê người phiên âm văn bản viết từ video,…

3.2. Tiến hành thu thập ngữ liệu

Đối mặt với vấn đề chi phí cũng như thời gian xây dựng, chúng tôi tập trung thu thập ngữ liệu bằng cách:

- Văn bản viết: thu thập những văn bản đã có bản điện tử (file word, ebook, bài viết trên website,…)
- Văn bản nói: thu thập những văn bản nói đã được phiên âm từ phụ đề của các video, phim tiếng Việt, phim tiếng Anh, một phần phiên âm thủ công từ video để thu ngữ liệu nói sử dụng tại ba miền ở Việt Nam.

3.2.1. Xác định các tiêu chí chung

Lựa chọn các tiêu chí lấy mẫu trong quá trình xây dựng kho ngữ liệu đóng vai trò quan trọng và mang tính chất quyết định đến thời gian xây dựng kho ngữ liệu, chi phí thực hiện, độ phức tạp và khả năng sai lệch trong quá trình lấy mẫu ngôn ngữ [8, tr. 25]. Các tiêu chí chung đơn giản nên được ưu tiên lựa chọn trong giai đoạn đầu và loại bỏ những tiêu chí phức tạp để tiết kiệm thời gian cũng như nhân lực.

Khi xây dựng kho ngữ liệu VnC 1.0, chúng tôi lựa chọn những tiêu chí chung được trình bày tại bảng 1. Những tiêu chí này đã được lựa chọn để xây dựng các kho ngữ liệu trên thế giới [16, tr. 10] đồng thời được cân nhắc đến khả năng thu thập ngữ liệu trong điều kiện thực tế. Ví dụ, nếu xét về tiêu chí thu thập văn bản theo địa điểm. Nếu chúng tôi lựa chọn theo vùng địa lí tại Việt Nam thì độ khó khi thu thập ngữ liệu sẽ tăng lên và tốn thời gian tìm kiếm nguồn ngữ liệu (đặc biệt là vấn đề khan hiếm ngữ liệu nói).

Bảng 1. Tiêu chí chung khi xây dựng kho ngữ liệu VnC 1.0

STT	Tiêu chí chung	Tiêu chí chung thu thập ngữ liệu VnC 1.0
1	Dạng thức của văn bản	Văn bản viết (văn bản đã được số hoá), văn bản nói (văn bản đã được phiên âm từ các video, lời thoại trong kịch bản phim)
2	Thể loại của văn bản	Sách, quyết định/quy định, báo cáo, văn kiện, kịch bản, luận văn, báo điện tử,…
3	Lĩnh vực của văn bản	Chuyên ngành và phổ thông
4	Địa điểm văn bản được tạo ra	Địa điểm: ba miền Bắc, Trung, Nam tại Việt Nam
5	Ngôn ngữ	Đơn ngữ - tiếng Việt.
6	Thời điểm văn bản được tạo ra	Các văn bản được xuất bản: trước năm 1945; 1945 - 1975; 1975 - 2000; chủ yếu từ năm 2000 đến 2020.

3.2.2. Thu thập ngữ liệu

Ngoài việc xác định các tiêu chí khi thu thập ngữ liệu còn phải quan tâm đến tính đại diện, tính cân bằng cũng như kích thước kho ngữ liệu khi thu thập [16, tr. 12-14].

Về tính đại diện: chúng tôi lựa chọn ngữ liệu từ những nguồn có uy tín, độ phổ biến cao. Ví dụ văn bản nói trong cuộc họp quốc hội sẽ được thu thập trực tiếp từ trang web quochoi.vn; các văn bản viết về báo cáo thường niên của các công ti, chúng tôi thu thập trực tiếp từ trang web chính thức của các công ti; các văn bản cũng phải đại diện cho nhiều lĩnh vực, được viết/nói ở nhiều vùng/miền,…

Về tính cân bằng: chúng tôi lựa chọn ngữ liệu đảm bảo cân bằng giữa các yếu tố xã hội của tác giả (giới tính, tuổi,…); giữa những người tiếp nhận văn bản (công chúng, nhóm người, một số người thân quen, trang trọng hay thân mật,…); cân bằng về vùng miền, thể loại văn bản.

Về kích thước kho ngữ liệu: trên thực tế không có kích thước tối đa, kho ngữ liệu BNC đã được xây dựng với kích thước gồm 100 triệu từ, đây là kho ngữ liệu đảm bảo tính cân bằng và đại diện cho tiếng Anh (được người Anh sử dụng), được sử dụng để khai thác phục vụ cho nhiều nghiên cứu ngôn ngữ [14]. Dựa trên đặc điểm của kho ngữ liệu BNC, chúng tôi xác định kích thước kho ngữ liệu tiếng Việt là 100 triệu từ, với kích thước này kho ngữ liệu đủ lớn để có thể đại diện cho tiếng Việt, các kết quả thống kê từ kho ngữ liệu có ý nghĩa cũng như phù hợp với điều kiện xây dựng kho ngữ liệu của chúng tôi.

Các tiêu chí thu thập ngữ liệu được chia làm hai loại đó là tiêu chí ngoài (external criteria): những tiêu chí phi ngôn ngữ (giới tính, độ tuổi, nghề nghiệp tác giả,…) và tiêu chí trong (internal criteria): những tiêu chí cơ bản về ngôn ngữ (từ vựng, cú pháp,…). Khi bắt đầu xây dựng kho ngữ liệu chúng ta có thể lựa chọn mẫu văn bản chủ yếu theo tiêu chí ngoài [23, tr. 8]. Bảng 2 trình bày các trình tiêu chí ngoài mà chúng tôi lựa chọn để thu thập văn bản viết và văn bản nói trong kho ngữ liệu VnC 1.0.

Bảng 2. Các tiêu chí ngoài khi thu thập ngữ liệu

Tiêu chí ngoài khi thu thập	**Các tiêu chí**
Thu thập văn bản viết	
Loại văn bản	Văn bản viết; văn bản nói; văn bản viết để nói: kịch bản Trong đó văn bản viết chiếm 90% và văn bản nói chiếm 10%
Lĩnh vực của văn bản	- Văn bản viết thông tin: tài liệu chuyên ngành, báo cáo, văn kiện, báo điện tử, luận văn, thường thức phổ thông. - Văn bản viết hư cấu: thơ, truyện ngắn/tiểu thuyết, tiểu sử tác giả, hồi kí.
Số tác giả sáng tác văn bản	Một tác giả; nhiều tác giả; nhiều tác giả cùng viết một tác phẩm
Thể loại văn bản	Sách; báo điện tử; báo cáo; văn kiện Quốc hội; các văn bản khác
Giới tính tác giả	- Nam/nữ - Nhiều giới tính (trường hợp nhiều tác giả cùng viết một tác phẩm) - Không rõ giới tính (ví dụ bài viết trên báo điện tử)
Kiểu lấy mẫu	- Toàn văn bản (dùng chủ yếu khi thu thập bài viết trên báo điện tử) - Một phần văn bản (dùng chủ yếu dùng khi thu thập tiểu thuyết) - Không rõ
Thu thập văn bản nói	
Vùng/miền	Bắc - Trung - Nam
Kiểu giao tiếp	Đối thoại - độc thoại
Ngữ cảnh giao tiếp	Truyền đạt thông tin/giáo dục; kinh doanh; bài phát biểu; phỏng vấn; sở thích

3.3. Xử lí ngữ liệu

3.3.1. Tiền xử lí ngữ liệu

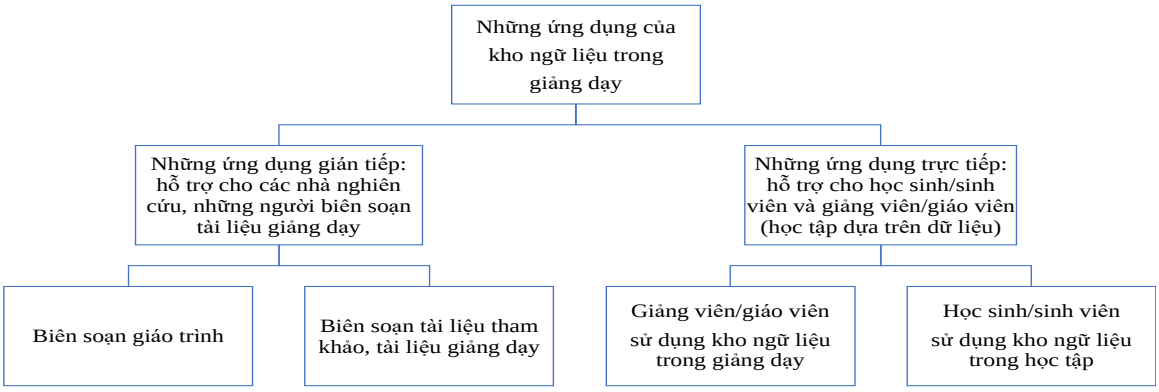
Sau khi thu thập ngữ liệu theo kế hoạch và dựa vào các tiêu chí, ngữ liệu cần được xử lí trước khi đưa vào gán nhãn thông tin ngôn ngữ. Do ngữ liệu được thu thập từ nhiều nguồn, chúng tôi tiến hành loại bỏ những thông tin không cần thiết như bảng biểu, sơ đồ, chú thích (chủ yếu trong các báo cáo thường niên); hình ảnh, video, các đường liên kết (link) (chủ yếu trong văn bản báo điện tử),... Sau đó chúng tôi tiến hành chuẩn hoá ngữ liệu (chuẩn hoá về dạng tập tin văn bản (.txt), bảng mã Unicode - utf8), kiểm lỗi ngữ liệu: các lỗi thường gặp phải như lỗi đánh máy, lỗi chính tả, lỗi về vị trí dấu thanh. Cuối cùng, chúng tôi phân đoạn câu trong từng văn bản để phục vụ cho các mục đích khai thác ngữ liệu.

3.3.2. Gán nhãn ngữ liệu

Để phục vụ dạy tiếng Việt cho người nước ngoài, kho ngữ liệu cần mang nhiều thông tin ngôn ngữ, vì thế đối với kho ngữ liệu VnC 1.0 chúng tôi gán nhãn ranh giới từ và gán nhãn từ loại tự động bằng công cụ gán nhãn do Trung tâm Ngôn ngữ học Tính toán xây dựng (CLC_VN_Tookit) [2].

4. Kho ngữ liệu phục vụ giảng dạy tiếng Việt cho người nước ngoài

Trên thế giới, kho ngữ liệu không chỉ phục vụ trực tiếp trong việc giảng dạy ngôn ngữ thứ hai mà còn hỗ trợ gián tiếp trong việc xây dựng giáo trình hay học liệu phục vụ dạy học (Hình 1). Đối với kho ngữ liệu VnC đã xây dựng, các kết quả rút trích từ kho ngữ liệu có nhiều ứng dụng phục vụ giảng dạy tiếng Việt cho người nước ngoài.



Hình 1. Ứng dụng của kho ngữ liệu trong dạy - học ngôn ngữ thứ hai [27, tr. 207]

4.1. Ứng dụng trong dạy từ vựng

Trong suốt quá trình học tập của người học, từ giai đoạn đầu [24, tr. 38] đến các giai đoạn sau [18, tr. 484], tần số từ vẫn đóng vai trò quan trọng, hỗ trợ người học học ngoại ngữ. Để xác định được vốn từ cơ bản phù hợp với người học ở giai đoạn đầu hay các vốn từ nâng cao ở các giai đoạn tiếp theo thì kết quả thống kê tần suất từ (Bảng 3) và kết quả thống kê tần suất tổng tích luỹ của từ

tiếng Việt (Hình 2) trong VnC 1.0 cho chúng ta xác định được số lượng từ cũng như mức độ thông dụng của chúng. Tần suất từ f được tính như sau:

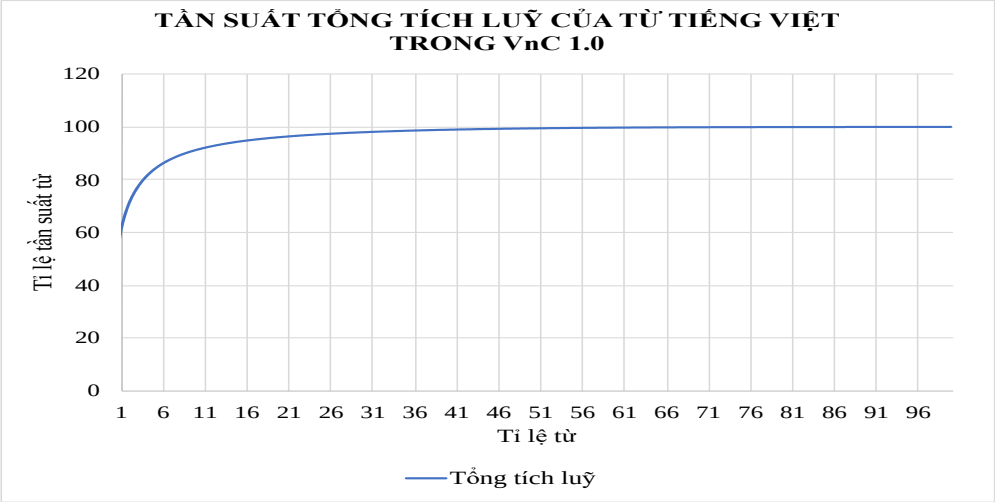
$$f = -\log\frac{n}{N}$$

Với f là tần suất từ (con số này càng nhỏ, từ này xuất hiện càng nhiều), n là số lần xuất hiện của từ trong kho ngữ liệu VnC 1.0, N là tổng số từ trong kho ngữ liệu VnC 1.0 (với N = 100.189.286 từ).

Bảng 3 trình bày một số kết quả thống kê tần suất từ trong VnC 1.0, đối với tần suất từ theo từ loại, kí hiệu Cm là giới từ, Cp: liên từ, Vc: hệ từ, R: phụ từ, Ve: động từ tồn tại, Pd: đại từ không gian, thời gian, Vv: động từ ý chí, Aa: tính từ hàm chất. Danh sách thống kê tần suất từ theo từ loại cung cấp cho chúng ta thông tin về nghĩa thông dụng của từ, từ đó lựa chọn giảng dạy từ có nghĩa phù hợp với trình độ của người học. Ví dụ từ *của* ở thể giới từ (f = 1,943) xuất hiện rất nhiều lần so với từ "của" ở thể danh từ (f = 4,510).

Bảng 3. Thống kê tần suất từ, tần suất từ theo từ loại trong kho ngữ liệu VnC 1.0

STT	Thống kê tần suất từ			Thống kê tần suất từ theo từ loại			
	Từ	f	n	Từ	POS	f	n
1	*và*	1,930	1.396.663	*của*	Cm	1,943	1.353.871
2	*của*	1,942	1.357.606	*và*	Cp	1,953	1.323.730
3	*là*	1,955	1.318.331	*là*	Vc	1,989	1.217.302
4	*có*	1,996	1.198.639	*không*	R	2,020	1.134.407
5	*không*	2,020	1.134.897	*có*	Ve	2,056	1.045.227
…							
20	*đến*	2,341	542.106	*đó*	Pd	2,376	499.808
50	*nó*	2,621	283.961	*theo*	Vv	2,663	258.179
100	*chị*	2,948	134.015	*rằng*	Cp	2,970	127.264
1000	*giường*	3,934	13.830	*dễ_dàng*	Aa	3,912	14.530



Hình 2. Thống kê tần suất tổng tích luỹ của từ tiếng Việt trong VnC 1.0

Từ kết quả thống kê tần suất tổng tích luỹ của từ tiếng Việt ta thấy 1% các từ phổ biến nhất (334 từ) chiếm hơn 60% số lượt từ xuất hiện trong VnC 1.0 và 10% các từ phổ biến nhất (3.334 từ) chiếm hơn 90% (91,356 %) số lượt từ sử dụng trong VnC 1.0. Dựa vào danh sách khoảng 3.300 từ thông dụng trong kho ngữ liệu này, chúng ta có thể xây dựng được vốn từ phù hợp theo trình độ để người học có thể đọc hiểu được văn bản tiếng Việt. Bên cạnh đó, từ các kết quả thống kê, dựa vào từ điển tần số chúng ta có thể xây dựng các danh sách từ vựng khác nhau theo trình độ của học viên (như danh sách 1.000 từ phổ biến, 2.000 từ phổ biến, 3.000 từ phổ biến,…)

4.2. Biên soạn giáo trình theo từng cấp độ

Việc biên soạn giáo trình cho người học nhất thiết phải quan tâm đến độ khó văn bản nhằm lựa chọn bài đọc, bài tập, vốn từ phù hợp cho từng học viên ở từng cấp độ [8, tr. 262]. Từ kho ngữ liệu VnC 1.0 có thể rút trích nhiều bài đọc thuộc nhiều chủ đề (văn hoá, giáo dục, pháp luật, thể thao,…) và từ thể loại văn bản khác nhau (sách, báo, báo cáo,…) theo từng cấp độ (dễ, trung bình, khó) để phù hợp với trình độ của người học, phù hợp với mạch bài đọc của giáo trình. Giúp cho việc biên soạn giáo trình tránh được vấn đề biên soạn theo cảm tính và chủ quan.

4.3. Khảo sát chuỗi đồng hiện

Khi cần đưa ví dụ về cách dùng từ, khảo sát sự kết hợp của các từ, chúng ta có thể dựa vào kết quả khai thác chuỗi đồng hiện từ kho ngữ liệu (concordance - một số từ thường cùng xuất hiện chung với nhau). Những ví dụ được rút trích từ kho ngữ liệu là những ví dụ được sử dụng trong văn nói/văn viết trong thực tế. Ví dụ: khi đưa ra ví dụ về cách đặt câu với từ "gió chướng" cũng như khảo sát những từ kết hợp với từ này, khi khảo sát trên kho ngữ liệu VnC 1.0 chúng ta sẽ được kết quả như Hình 3. Người học có thể nhận thấy các từ như: *theo*, *luồng*, *chờ*, *mùa*,… có thể đứng trước từ "gió chướng".

Cha quấn khăn rằn theo	gió_chướng	
Ở đó có những ngày	gió_chướng	bông so_đũa nở trắng , cá_lòng_tong lũ_lượt kéo về từng bầy .
Trong luồng	gió_chướng	người_ta ngửi được cả mùi nôn_nao tết .
Tháng chạp	gió_chướng	nôn_nao e_ấp thập_thò ngoài ngõ .
Thời_điểm mùa khô 2014 , ở ven biển Gò_Công cũng trùng với mùa	gió_chướng	thổi mạnh nên tốc_độ xâm_thực của biển khiến đai rừng ngoài đê
Nhưng chờ	gió_chướng	lùa nắng về lại Nhơn_Thành thì cũng phải vài_ba tháng nữa .
Cuộc_sống quanh_năm chỉ trông_chờ vào mùa lúa , đó là mùa	gió_chướng	mà người Bạc_Liêu gọi là mùa vui .
Thà trở_về đúng vào mùa	gió_chướng	, mới hay má nó đã qua đời .

Hình 3. Chuỗi đồng hiện từ "gió chướng" được rút trích từ kho ngữ liệu.

4.4. Rút trích kết ngôn

Người nước ngoài rất khó hiểu và khó nhớ những cụm từ mà người Việt Nam thường dùng chung với nhau, ví dụ như: *ngựa ô/~~chó ô~~, ăn kiêng/~~uống kiêng~~, gà trống/~~đực~~*,… Vì thế, khi rút trích các chuỗi kết ngôn này từ kho ngữ liệu VnC 1.0 (Hình 4), người học có thể biết cách dùng từ tích hợp trong ngữ cảnh thích hợp [8, tr. 262].

Chế_độ	ăn kiêng	không có các thực_phẩm từ thịt , kém hấp_thu ruột , rối_loạn chức_n
- Chính vì vậy , cả bệnh_lý ở da và ở ruột của người bệnh đều giảm khi chế_độ	ăn kiêng	không có gluten và nặng lên khi người bệnh ăn chế_độ có gluten .
Đó chính là điều tôi đang nói đến , " tôi đáp " Điều đó cũng giải_thích tại_sao mà những chương_trình tập_luyện và	ăn kiêng	hiệu_quả nhất lại không thành_công nếu_như một người không thực_
Một ví_dụ đơn_sơ cho chuyện này là trường_hợp một người muốn bắt_đầu	ăn kiêng	, nhưng cảm_thấy chần_chờ , chủ_yếu là vì nơi tiềm_thức họ , họ tin
Tại_sao bạn không bắt_đầu	ăn kiêng	?
Nếu tôi	ăn kiêng	, chịu đói , để rồi rốt_cuộc vẫn lên_cân thì sao ?
Cách tốt nhất là theo " chế_độ	ăn kiêng	tinh_thần " .
Chế_độ	ăn kiêng	tinh_thần là cơ_hội để bạn loại_bỏ những thói_quen tiêu_cực tai_hại
Tôi đề_nghị với bạn chế_độ	ăn kiêng	tinh_thần sau đây : Trong mười ngày , bắt_đầu ngay từ hôm_nay , bạ
Nhưng những người đã từng bắt_đầu áp_dụng chế_độ	ăn kiêng	tinh_thần này đều phải ngạc_nhiên khi khám_phá ra rằng tâm_trí họ

Hình 4. Ví dụ về kết ngôn "ăn kiêng" được rút trích tự động từ kho ngữ liệu VnC 1.0

5. Kết luận

Trong bài báo này chúng tôi đã trình bày cách thức xây dựng kho ngữ liệu tiếng Việt VnC 1.0 phục vụ việc dạy tiếng Việt cho người nước ngoài. Kho ngữ liệu này có kích thước lớn, được xây dựng cân bằng và đảm bảo tính đại diện cho tiếng Việt. Các kết quả khai thác từ kho ngữ liệu này ứng dụng vào dạy tiếng Việt cho người nước ngoài một cách trực tiếp và gián tiếp trong các bước: dạy từ vựng, biên soạn giáo trình, bài tập, dạy học dựa vào chuỗi đồng hiện. Trong thời gian tới, chúng tôi tiếp tục phát triển kho ngữ liệu, tăng số lượng mẫu văn bản, điều chỉnh tỉ lệ giữa các mẫu văn bản cũng như gán thêm thông tin ngôn ngữ để các kết quả rút trích từ kho ngữ liệu mang nhiều tính ứng dụng hơn.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Bạch Thanh Minh, *Giảng dạy Tiếng Việt cho người nước ngoài nói tiếng Anh bằng phương pháp giao tiếp*, Luận văn Thạc sĩ Việt Nam học, Đại học Quốc tế Hồng Bàng, Thành phố Hồ Chí Minh, 2016.
2. CLC Tagging Toolkits - *Công cụ gán nhãn ngữ liệu do Trung tâm Ngôn ngữ học Tính toán xây dựng*, Truy cập 10/08/2020 tại http://nlpusr.kimtudien.com.vn/
3. Dương Thị Thu, *Xử lí ngữ liệu thành ngữ tục ngữ trong giảng dạy tiếng Việt cho người nước ngoài*, Luận văn Thạc sĩ Việt Nam học, Đại học Khoa học xã hội và Nhân văn - Đại học Quốc gia Hà Nội, 2018.
4. Đào Hồng Thu, *Sử dụng khối tư liệu giảng dạy tiếng Việt như một ngoại ngữ,* Hội thảo Quốc tế *Nghiên cứu và giảng dạy ngôn ngữ, văn hóa Việt Nam - Trung Quốc ở Đông Á và Đông Nam Á* lần thứ II, Hà Nội, 2009.
5. Đinh Điền và Hồ Xuân Vinh, *Ứng dụng Kho Ngữ liệu trong việc day tiếng Việt cho người nước ngoài*, Hội Thảo Quốc tế Việt Nam học, Bình Châu - Bà Rịa - Vũng Tàu, 2016.
6. Đinh Điền và Lý Ngọc Minh, *Ứng dụng Ngữ liệu Song ngữ Anh-Việt trong Giảng dạy Ngôn ngữ*, Hội Thảo Liên ngành Ngôn ngữ học ứng dụng & Giảng dạy ngôn ngữ, Huế, 2015.
7. Đinh Điền, *Ứng dụng ngôn ngữ học tính toán trong nghiên cứu và giảng dạy ngôn ngữ*, Hội thảo Khoa học Quốc gia, Hà Nội, 2019.
8. Đinh Điền, *Ngôn ngữ học ngữ liệu*, Nxb Đại học Quốc gia Thành phố Hồ Chí Minh, 2018.
9. Phạm Hiển, *Sử dụng kho ngữ liệu trong giảng dạy tiếng Việt*, Từ điển học và Bách khoa thư, 2012.
10. Vũ Lan Hương, *Giảng dạy hành động ngôn từ tiếng Việt cho người nước ngoài theo định hướng chức năng và góc độ giao tiếp liên văn hóa - VSL*, Ngôn ngữ & Đời sống, 2018.

Tiếng Anh

11. Ewa Donesch-Jezo, *Using language corpus in teaching foreign language vocabulary*, International Multidisciplinary e-Journal, India, 2013.
12. Gena R. Bennett, *Using corpora in the language learning classroom: Corpus linguistics for teachers*, University of Michigan Press, United States, 2010.
13. Kiyomi Chujo, Masao Utiyama, and Shinji Miura, *Using a Japanese - English parallel corpus for teaching English vocabulary to beginning-level students*, English Corpus Studies, 2006.
14. Lou Burnard (Edited), *Reference Guide for the British National Corpus (XML Edition)*, Oxford University Computing Services, 2007, Retrieved February 23, 2020 from http://www.natcorp.ox.ac.uk/docs/URG/
15. Mark Davies, *The Corpus of contemporary American English as the first reliable monitor corpus of English*, Literary and Linguistic Computing, United Kingdom, 2010.
16. Martin Wynne (Edited), *Developing linguistic corpora: A guide to good practice (AHDS guides to good practice)*, Oxbow Books, United Kingdom, 2005.
17. Michael McCarthy, Jeanne McCarten and Helen Sandiford, *Touchstone 4 (Teacher's Edition)*, Cambridge University Press, United Kingdom, 2006.
18. Norbert Schmitt and Diane Schmitt, *A reassessment of frequency and vocabulary size in L2 vocabulary teaching*, Cambridge University Press, United Kingdom, 2012.
19. Paul Thompson, *Review of exploring Academic English: A workbook for student essay writing*, Language Learning & Technology, United States, 2001.
20. Randi Reppen, *Using corpora in the language classroom*, Cambridge University Press, United Kingdom, 2010.
21. Scott Thornbury, *Natural grammar: The keywords of English and how they work*, OUP Oxford, United Kingdom, 2004.
22. SIL, *Language Report*, 2017, Retrieved September 25, 2020, from http://www.ethnologue.com
23. Sue Atkins, Jeremy Clear and Nicholas Ostler, *Corpus design criteria*, Literary and 24, Linguistic Computing, United Kingdom, 1992.
24. Tom Cobb, *Computing the vocabulary demands of L2 reading*, Language Learning and Technology, United States, 2007.
25. Tony McEnery, Richard Xiao and Yukio Tono, *Corpus survey*, Lancaster University, 2013, Retrieved September 5, 2020, from https://www.lancaster.ac.uk/fass/projects/corpus/cbls/corpora.asp#_Toc92298862
26. Totok Suhardijanto and Deodatus Perdana Putra, *Acquiring extended units of meaning: The Role of learner corpus in teaching Indonesian as a foreign language*, Proceedings of the 2nd Konferensi BIPA Tahunan by Postgraduate Program of Javanese Literature and Language Education in Collaboration with Association of Indonesian Language and Literature Lecturers, Indonesia, 2019.
27. Ute Römer, *Annual review of applied linguistics: Corpus research applications in second language teaching*, Cambridge University Press, United Kingdom, 2011.
28. Vi Thanh Son, *Using corpus linguistics in teaching and learning English in the classroom (For Vietnamese learners of English)*, Lund University Publications, Sweden, 2014, Retrieved September 20, 2020 from http://lup.lub.lu.se/record/4589482