NGÔN NGỮ HỌC KHỐI LIỆU – KHÁI NIỆM, CÁCH TIẾP CẬN, PHƯƠNG PHÁP VÀ ỨNG DỤNG TRONG NGHIÊN CỨU, GIẢNG DẠY TIẾNG ĐỨC NHƯ MỘT NGOẠI NGỮ

Lê Tuyết Nga*

Khoa Ngôn ngữ và Văn hóa Đức, Trường Đại học Ngoại ngữ, ĐHQGHN, Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Nhận bài ngày 24 tháng 7 năm 2020
Chỉnh sửa ngày 27 tháng 8 năm 2020; Chấp nhận ngày 15 tháng 9 năm 2020

Tóm tắt: Bài viết[1] bàn thảo về khái niệm khối liệu (định nghĩa, các tiêu chí xác định khối liệu, phân loại khối liệu), ngôn ngữ học khối liệu như một ngành khoa học hoặc như một phương pháp luận, các cách tiếp cận (cách tiếp cận dựa vào khối liệu để kiểm chứng lí thuyết và cách tiếp cận được chỉ dẫn bởi khối liệu để xây dựng lí thuyết), các phương pháp nghiên cứu (định lượng, định tính) cũng như các công cụ được sử dụng trong ngôn ngữ học khối liệu nhìn từ góc độ của các nhà khoa học Đức. Một trọng tâm của bài viết là mối liên hệ giữa ngôn ngữ học khối liệu và việc giảng dạy tiếng Đức như một ngoại ngữ, những khả năng ứng dụng của ngôn ngữ học khối liệu vào nghiên cứu và giảng dạy tiếng Đức.

Từ khóa: khối liệu, ngôn ngữ học khối liệu, cách tiếp cận, phương pháp, tiếng Đức như một ngoại ngữ

1. Đặt vấn đề

Trong nghiên cứu và giảng dạy ngôn ngữ nói chung và tiếng Đức nói riêng, ta thường gặp phải những tình huống sau đây:

(a) Nên chọn từ nào hoặc cách diễn đạt nào, ví dụ "Wie lösen wir dieses schwere/schwierige Problem?" (Andresen và Zinsmeister, 2019, tr. 1) hoặc "wegen des schlechten Wetters" (cách 2/ sở hữu cách) hay "wegen dem schlechten Wetter" (cách 3/tặng cách)? Một trong nhiều cách để tìm lời giải đáp cho những câu hỏi này là nghiên cứu tần số xuất hiện của các cách sử dụng những từ và diễn đạt này trong thực tế nhờ các khối liệu (corpus) điện tử. Theo một nghiên cứu về việc sử dụng *wegen* (*vì*) ở khoảng 200 tờ báo tiếng Đức trong thời gian 5 tuần của Elter (2005) (dẫn theo Scherer, 2014, tr. 3), trung bình mỗi ngày *wegen* xuất hiện 299 lượt ở cách 2 và chỉ có 2,5 lượt ở cách 3. Như vậy với khối liệu này, Elter có thể chứng minh rằng ở văn phong báo chí thì *wegen* hầu như chỉ được sử dụng ở cách 2.

(b) Khi lựa chọn những hiện tượng ngữ pháp cần được đưa vào giáo trình giảng dạy thì một trong những tiêu chí được sử dụng là tần số xuất hiện của chúng trong các văn bản. Ví dụ theo Jones và Tschirner (2006) và Tschirner (2008) thì những giới từ sau xuất hiện trong 20 từ có tần số cao nhất: in (4), zu (6), von (11), mit (13), auf (17), für (18), an (19). Còn theo khối liệu Duden[2], trong 17,4

* ĐT: 84-904108681
Email: ngatoan@gmail.com

[1] Nghiên cứu này được hoàn thành với sự hỗ trợ của Trường Đại học Ngoại ngữ, Đại học Quốc gia Hà Nội trong đề tài mã số N.19.05

[2] Truy cập lúc 11:00 ngày 17/7/2020 tại https://www.duden.de/sprachwissen/sprachratgeber/Die-haufigsten-Worter-deutschsprachigen-Texten

triệu từ gốc thì các giới từ trên xếp hạng như sau: in (2), zu (6), von (7), mit (10), an (11), für (12), auf (13). Chúng ta có thể dễ dàng tìm thấy tất cả các giới từ này trong bảng tổng hợp ngữ pháp của các giáo trình tiếng Đức trình độ A1.

(c) Để đưa ra các biện pháp cải tiến phương pháp và học liệu giảng dạy, thông thường chúng ta dựa vào kinh nghiệm giảng dạy, quan sát và theo dõi quá trình học tập, sử dụng những hiểu biết về tiếng mẹ đẻ và ngoại ngữ để đưa ra các giả thuyết về những vấn đề của người học cần được khắc phục. Tuy nhiên những giả thuyết này vẫn cần phải được kiểm chứng thông qua những kết quả nghiên cứu thực nghiệm đáng tin cậy về năng lực làm chủ ngôn ngữ thực tế của người học. Những nghiên cứu này chỉ có thể thực hiện được dựa trên phân tích những ngữ liệu xác thực trong một khối liệu người học cụ thể.

Những ví dụ trên cho thấy nhiều câu hỏi nghiên cứu và ứng dụng có thể được giải quyết nhờ các nghiên cứu thực nghiệm một cách hệ thống dựa vào các khối liệu ngôn ngữ (linguistic corpus). So với tra cứu trên internet, google thì ưu điểm nổi trội của các khối liệu này là nội dung của chúng xác thực, có thể được kiểm chứng, không bị tác động bởi những thay đổi thường xuyên đồng thời những thông tin về nguồn gốc, số lượng, thời gian v.v. vào thời điểm truy cập là chính xác (Andresen và Zinsmeister, 2019, tr. 9). Vì vậy có thể nói việc sử dụng khối liệu để tìm các giải pháp cho nghiên cứu và giảng dạy ngôn ngữ đang nhận được sự quan tâm của nhiều nhà khoa học.

Mục tiêu của bài viết này là đưa ra cái nhìn khái quát về ngôn ngữ học khối liệu ở Đức và từ góc độ của các nhà nghiên cứu Đức cùng các cách tiếp cận, phương pháp và công cụ nghiên cứu, ứng dụng trong nghiên cứu và giảng dạy tiếng Đức, từ đó đưa ra một số đề xuất cho việc phát triển ngôn ngữ học khối liệu ở Đức cũng như ở Việt Nam và khu vực.

2. Khối liệu và ngôn ngữ học khối liệu

2.1. Ngôn ngữ học khối liệu

Trong khi ngôn ngữ học khối liệu (corpus linguistics) như một phân ngành ngôn ngữ trong nghiên cứu tiếng Anh đã hình thành và phát triển từ thập kỉ 90 của thế kỉ trước thì ngành ngôn ngữ Đức và chuyên ngành Tiếng Đức như một ngoại ngữ mới bắt đầu sử dụng các phương pháp của ngôn ngữ học khối liệu để giải quyết các câu hỏi nghiên cứu từ đầu thế kỉ 21 (Fandrych và Tschirner, 2007, tr. 195). Những dẫn luận đầu tiên và khái quát về ngôn ngữ học khối liệu xuất hiện vào năm 2006 với các tác giả Lemnitzer và Zinsmeister cũng như Scherer, tiếp theo đó là các nghiên cứu của Lüdeling và Walter (2010a), Keibel và cộng sự (2012), Kupietz và Schmidt (2018), Andresen và Zinsmeister (2019), Hirschmann (2019). Trong những tác giả viết về mối liên hệ giữa ngôn ngữ học khối liệu và nghiên cứu, giảng dạy ngoại ngữ cũng như nghiên cứu quá trình thụ đắc ngoại ngữ thì phải kể đến Fandrych và Tschirner (2007), Lüdeling và cộng sự (2008), Tschirner (2009), Lüdeling và Walter (2010b), Ahrenholz và Wallner (2013), Fandrych và cộng sự (2018).

Trong các nghiên cứu này, các nhà ngôn ngữ học đưa ra hai cách hiểu về ngôn ngữ học khối liệu. Scherer (2014) và Hirschmann (2019) định nghĩa ngôn ngữ học khối liệu từ góc độ phương pháp luận. Theo đó ngôn ngữ

học khối liệu là "một trong những phương pháp nhằm nghiên cứu việc sử dụng ngôn ngữ thông qua các dữ liệu xác thực" (Scherer, 2014, tr. 2) hoặc là "một phương pháp nghiên cứu thực nghiệm với mục tiêu giải quyết các câu hỏi nghiên cứu trong ngôn ngữ học" nhờ các dữ liệu được thu thập đáp ứng mục tiêu nghiên cứu (Hirschmann, 2019, tr. 1). Keibel và cộng sự (2012, tr. 20-21) quan niệm ngôn ngữ học như một phương pháp luận (corpus linguistics as a methodology), không phải là một hộp công cụ (tool box), với trọng tâm không phải là dựa vào khối liệu (corpus-based) để khẳng định hoặc phản bác các giả thuyết hoặc lý thuyết, mà khối liệu được coi là điểm khởi đầu của nghiên cứu. Các nhà nghiên cứu không đưa ra các giả thuyết, thay vì đó họ hoàn toàn định hướng vào việc sử dụng ngôn ngữ trong thực tế, tìm ra các qui luật và xây dựng lý thuyết, giả thuyết từ kết quả nghiên cứu dữ liệu (corpus-driven) (cụ thể xem thêm mục 3.1).

Theo Lemnitzer và Zinsmeister (2015, tr. 14-15) thì ngôn ngữ học khối liệu là ngành khoa học mô tả "các phát ngôn của ngôn ngữ tự nhiên, các thành tố và cấu trúc của chúng" và xây dựng cơ sở lý luận "dựa trên nền tảng phân tích các văn bản xác thực được tập hợp thành một khối liệu". Là một ngành khoa học nên ngôn ngữ học khối liệu "phải tuân theo các nguyên tắc khoa học và đáp ứng các yêu cầu về khoa học". Kết quả các nghiên cứu có thể phục vụ cho việc giảng dạy ngoại ngữ, cung cấp các tư liệu về ngôn ngữ, xử lý dữ liệu ngôn ngữ điện tử, từ điển học và ngôn ngữ học máy tính (dẫn theo Lê Tuyết Nga, 2020, tr. 353). Đối tượng nghiên cứu của ngôn ngữ học khối liệu theo Lüdeling và Walter (2010a, tr. 315) là quá trình xây dựng khối liệu, cấu trúc khối liệu, chú giải ngôn ngữ và siêu ngôn ngữ cũng như xử lý dữ liệu và sử dụng khối liệu. Tschirner (2009, tr. 127) và Lemnitzer và Zinsmeister (2015, tr. 11-12, 19-23) nêu bật điểm mạnh của ngôn ngữ học khối liệu (thuộc chủ nghĩa kinh nghiệm/chủ nghĩa duy nghiệm (empirism)) trong so sánh với ngữ pháp sản sinh (thuộc chủ nghĩa duy lý (rationalism)). Mục tiêu của ngữ pháp sản sinh là mô tả và giải thích năng lực ngôn ngữ (competence) trên cơ sở diễn giải duy lý và những đánh giá về năng lực ngữ pháp dựa trên những câu ví dụ do chính nhà nghiên cứu tạo ra và không gắn với một ngữ cảnh nào đó. Trái lại ngôn ngữ học khối liệu quan tâm tới các dữ liệu và ngữ liệu xác thực có thể quan sát được với mục tiêu mô tả và giải thích năng lực sử dụng ngôn ngữ (performance) nhờ vào việc phân tích một lượng văn bản lớn với sự hỗ trợ của công nghệ máy tính. Tuy nhấn mạnh tính xác thực của khối liệu nhưng Lemnitzer và Zinsmeister (2015, tr. 28-29) cũng nhận thấy một số vấn đề của khối liệu như sau: kích cỡ của khối liệu không rõ ràng và có thể không đủ để đại diện cho một ngôn ngữ; trong khối liệu xuất hiện những dữ liệu không quan trọng hoặc không liên quan; có những cấu trúc đúng ngữ pháp nhưng không xuất hiện trong khối liệu; trong khối liệu có những cấu trúc lệch chuẩn, không đúng ngữ pháp và do đó không đáng tin cậy.

Bên cạnh việc bàn thảo về quá trình phát triển của ngôn ngữ học khối liệu từ phương pháp luận thành một phân ngành khoa học trong ngôn ngữ học ứng dụng và giữ một "vị trí lịch sử" trong thời hiện đại, Klein (2013, tr. 336-340) đưa ra khái niệm "ngôn ngữ học ngân hàng dữ liệu" (data bank linguistics) như là sự phát triển tiếp theo của ngôn ngữ học khối liệu. Ngôn ngữ học ngân hàng dữ

liệu là một hình thức đặc biệt của ngôn ngữ học, trong đó việc sử dụng ngân hàng dữ liệu máy tính sẽ đóng vai trò then chốt trong nghiên cứu lý luận, thực tiễn và phương pháp luận. Điểm đặc biệt của việc ứng dụng các ngân hàng dữ liệu nằm ở ba lĩnh vực: nghiên cứu (mở rộng phạm vi cho các câu hỏi nghiên cứu); xử lý dữ liệu để truy cập được nhanh, dễ dàng và hệ thống; các kỹ thuật hỗ trợ mới (ví dụ để tìm ra các thông tin có tính hệ thống về siêu dữ liệu từ các bảng hỏi, phỏng vấn, các thí nghiệm hay văn bản). Klein (2013, tr. 340) cho rằng sử dụng ngân hàng dữ liệu không có nghĩa là ngay lập tức sẽ tạo ra một sự chuyển biến về chất mà thông qua việc mở rộng và hệ thống hóa các dữ liệu nhờ vào các khả năng mới của kỹ thuật máy tính - tức là tăng về lượng - ngôn ngữ học ngân hàng dữ liệu có thể nâng tiềm năng nhận thức lên một tầm cao mới. Tuy nhiên khái niệm này hiện vẫn chưa nhận được sự quan tâm của các nhà khoa học khác.

2.2. Khối liệu

Khái niệm "khối liệu" được dùng để chỉ một tập hợp văn bản hoặc trích đoạn văn bản xác thực trong ngôn ngữ viết và ngôn ngữ nói, được sản sinh trong ngữ cảnh cụ thể, được số hóa và có thể tìm kiếm bằng các công cụ điện tử (Lüdeling và Walter, 2010a, tr. 315; Lemnitzer và Zinsmeister, 2015, tr. 13; Meißner và cộng sự, 2016, tr. 307; Hirschmann, 2019, tr. 2). Khối liệu được xây dựng nhằm mục đích phục vụ cho các nghiên cứu thực nghiệm và đặc biệt hữu ích nếu bao gồm một lượng dữ liệu lớn được xử lý nhờ công nghệ máy tính. Đặc biệt quan trọng đối với lĩnh vực nghiên cứu giảng dạy và thụ đắc ngoại ngữ là khối liệu người học (learner corpus) thường bao gồm ngữ liệu của người học ngoại ngữ (như khối liệu GeWiss), có thể kèm theo phân loại lỗi và đưa ra giả thuyết chữa lỗi (như khối liệu Falko).

Bên cạnh các tiêu chí bắt buộc (dữ liệu có nguồn gốc và nội dung có thể kiểm chứng, được sản sinh trong bối cảnh ngôn ngữ tự nhiên và xác thực, ở dạng điện tử và có thể xử lý nhờ kĩ thuật máy tính) thì khối liệu còn đáp ứng các tiêu chí hoặc đặc trưng sau đây: (a) tính điển hình/tính đại diện, (b) sự gán nhãn siêu ngôn ngữ (metadata), (c) tính chú giải ngôn ngữ (annotation) (Keibel và cộng sự, 2012, tr. 57-59; Scherer, 2014, tr. 5-10; Lemnitzer và Zinsmeister, 2015, tr. 39-88; Hirschmann, 2019, tr. 2-4)[1]. Như vậy một khối liệu thường bao gồm ba loại dữ liệu: dữ liệu gốc, siêu dữ liệu và chú giải ngôn ngữ. Dữ liệu gốc (primary data) là các văn bản được tập hợp trong khối liệu và thường kèm theo các bản phiên âm đối với khối liệu ngôn ngữ nói. Hirschmann (2019, tr. 5-6) phân biệt ba nhóm dữ liệu gốc: Nhóm 1 (not elicited data) bao gồm các dữ liệu đã tồn tại và được sản sinh trong những ngữ cảnh xác thực như các văn bản trên diễn đàn internet, trên báo, tiểu thuyết, thư từ (ví dụ khối liệu TIGER[2]). Dữ liệu thuộc nhóm 2 (elicited data) được "thu thập cho một mục tiêu nghiên cứu nhất định" như các cuộc hội thoại trong những ngữ cảnh nhất định (ví dụ khối liệu FOLK[3]) hoặc bài viết, kết quả của các bảng hỏi. Nhóm 3 (experimental data) gồm những dữ liệu tương tự như nhóm 2 nhưng quá trình sản sinh và thu thập được giám sát một cách chặt chẽ, qua đó có thể "tác động lên những biến số nhất định

1 Xem thêm Lê Tuyết Nga, 2020, tr. 354-355.

2 Truy cập lúc 14:00 ngày 18.7.2020 tại https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger/

3 Truy cập lúc 14:20 ngày 18.7.2020 tại https://dig-hum.de/forschung/projekt/forschungs-und-lehrkorpus-gesprochenes-deutsch

nhằm kiểm chứng một giả thuyết hoặc trả lời cho một câu hỏi nghiên cứu" (ví dụ khối liệu ALC[1]). Siêu dữ liệu (metadata) được dùng để chỉ những "dữ liệu về dữ liệu" (Lemnitzer và Zinsmeister, 2015, tr. 44), những thông tin bổ sung liên quan đến dữ liệu gốc như tác giả, năm xuất bản, bối cảnh sản sinh văn bản, bối cảnh xuất bản, người thu thập dữ liệu, người xây dựng khối liệu, dữ liệu về người học (nằm trong khối liệu người học) và đặc biệt là thể loại văn bản. Dữ liệu chú giải ngôn ngữ bao gồm các phân tích dữ liệu gốc theo các phạm trù ngôn ngữ trên các bình diện hình thái, cú pháp, ngữ nghĩa, ngữ dụng và cấu trúc văn bản, ngoài ra còn có chú giải lỗi trong các khối liệu người học.

Ngoài phân loại khối liệu theo các tiêu chí như chức năng và mục đích sử dụng, phương tiện ngôn ngữ, độ lớn, tính chú giải, tính ổn định, lĩnh vực ứng dụng và tính sử dụng (Scherer, 2014; Lemnitzer và Zinsmeister 2015[2]), Fandrych và Tschirner (2007, tr. 202) còn phân biệt khối liệu bản ngữ (native corpus), khối liệu ngôn ngữ đặc dụng trong lớp học (classroom corpus) và khối liệu người học. Khối liệu bản ngữ với ngữ cảnh giao tiếp tự nhiên của người bản ngữ được xây dựng "nhằm phục vụ cho việc xác định nội dung học cũng như biên soạn học liệu xác thực" cho việc giảng dạy ngoại ngữ và có thể được sử dụng như một "khối liệu so sánh để nghiên cứu quá trình thụ đắc ngôn ngữ thứ hai" (Paschke, 2018, tr. 22). Khối liệu ngôn ngữ đặc dụng trong lớp học bao gồm các dữ liệu ở dạng video và audio, chủ yếu là các bài giảng và giờ học cũng như các bản phiên âm, học liệu và bản trình bày PowerPoint kèm theo. Có thể kể đến 3 khối liệu trong Dự án nghiên cứu quốc tế EuroWiss[3] gồm khoảng 350 giờ giảng với mục đích nghiên cứu phân tích diễn ngôn và so sánh phương pháp giảng dạy đại học. Một ví dụ khác là khối liệu ngôn ngữ đặc dụng trong lớp học tiếng Anh FLECC (The Flensburg English Classroom Corpus) với dữ liệu audio và phiên âm của 39 giờ học tiếng Anh từ lớp 3 đến lớp 10 tại các trường phổ thông ở bang Schleswig-Holstein (Bắc Đức) (Jäkel, 2010, tr. 9). Khối liệu này có thể được sử dụng như một học liệu đặc biệt hữu ích cho chương trình đào tạo giáo viên tiếng Anh hoặc để nghiên cứu phương pháp giảng dạy, lỗi và sự lệch chuẩn, tác phong và thái độ của giáo viên. Khối liệu người học là cơ sở để nghiên cứu lỗi, sự lệch chuẩn và quá trình thụ đắc ngoại ngữ. Hai khối liệu người học lớn nhất, trực tuyến và truy cập miễn phí là Falko[4] (gồm nhiều tiểu khối liệu như khối liệu bài viết của người học, khối liệu so sánh, khối liệu cắt dọc[5] v.v.) và Merlin[6] (gồm 2.286 văn bản viết của người học tiếng Đức, tiếng Ý và tiếng Tiệp được chú giải ở nhiều bình diện)[7].

1 Truy cập lúc 15:00 ngày 18.7.2020 tại https://www.phonetik.uni-muenchen.de/forschung/abgeschlossene_projekte/alc.html

2 Xem thêm Lê Tuyết Nga, 2020, tr. 355.

3 Truy cập lúc 15:07 ngày 18.7.2020 tại https://www.slm.uni-hamburg.de/forschung/forschungsprojekte/eurowiss.html

4 Truy cập lúc 15:58 ngày 18.7.2020 tại https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/design

5 Khối liệu cắt dọc (longitudinal corpus) bao gồm dữ liệu người học ở những thời điểm khác nhau để đánh giá sự tiến bộ của người học.

6 Truy cập lúc 16:00 ngày 18.7.2020 tại https://merlin-platform.eu/

7 Xem thêm Lê Tuyết Nga, 2020, tr. 356.

3. Các cách tiếp cận và các phương pháp nghiên cứu

3.1. Các cách tiếp cận

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). Bên cạnh thuật ngữ trong tiếng Anh, các nhà ngôn ngữ học Đức dùng khá nhiều thuật ngữ trong tiếng Đức để chỉ hai hướng tiếp cận này, chúng ta có thể xem bảng sau:

Bảng 1: Thuật ngữ chỉ cách tiếp cận trong ngôn ngữ học khối liệu (Keibel và cộng sự, 2012, tr. 19-21; Meißner, 2014, tr. 89-91; Lemnitzer và Zinsmeister, 2015, tr. 33-38; Brommer, 2018, tr. 102-105)

	Bubenhofer (2009)	Keibel và cộng sự (2012)	Bubenhofer (2009) Steyer (2013)	Meißner (2014)	Lemnitzer và Zinsmeister (2015)
corpus-based	korpusgeleitet als Oberbegriff	corpus-based	korpusbasiert	korpusbasiert	korpusgestützt
corpus-driven		corpus-driven, struktur-entdeckende Verfahren	korpusgesteuert	korpusgesteuert, datengeleitet	korpusbasiert
				korpusillustriert, corpus-illustrated	

Điểm chung của tất cả các tác giả là 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. Ngoài ra, Meißner (2014: tr. 89) nhắc tới corpus-illustrated như một cách tiếp cận minh họa bằng khối liệu mà khi đó khối liệu chỉ đóng vai trò là một "tập hợp ví dụ" nhằm xác nhận sự tồn tại của một đơn vị, một từ hoặc một cấu trúc.

Cách tiếp cận dựa vào khối liệu coi các dữ liệu trong một khối liệu là nguồn minh chứng bổ sung cho các lý thuyết ngôn ngữ nhằm kiểm nghiệm, xác nhận hoặc phủ nhận các giả thuyết nhất định. Trọng tâm của nghiên cứu là "các bằng chứng thực nghiệm và xu hướng định lượng" (Meißner, 2014, tr. 90). Một ví dụ cho cách tiếp cận này là nghiên cứu của Niederhaus (2011). Để kiểm nghiệm xem "mức độ chuyên ngành của các văn bản trong giáo trình dạy nghề có phụ thuộc vào chuyên ngành không" (Niederhaus, 2011, tr. 213), tác giả đã sử dụng hai khối liệu về chăm sóc cơ thể và kĩ thuật điện tử, nghiên cứu tần suất của các cấu trúc chuyên ngành điển hình như từ ghép, câu phức, định ngữ, bị động và so sánh các kết quả thống kê với nhau.

Cách tiếp cận được chỉ dẫn bởi khối liệu trao sự ưu tiên cho khối liệu và đòi hỏi lượng dữ liệu lớn với mục đích "phát hiện ra các hiện tượng và các liên kết mới, trước đó chưa được biết tới" (Köhler, 2005, tr. 4, dẫn theo Keibel và cộng sự, 2012, tr. 20-21), tạo ra các phạm trù phân tích và đơn vị mô tả "từ kết quả xử lý dữ liệu", qua đó tránh được trường

hợp chỉ nắm bắt các cấu trúc theo các phạm trù phân tích đã xác định trước đó (Meißner, 2014, tr. 90). Trong cách tiếp cận này, Lemnitzer và Zinsmeister (2015, tr. 33-38) phân biệt cách tiếp cận định lượng (corpus-driven quantitative approach) và cách tiếp cận định tính và định lượng (corpus-driven quantitative-qualitative approach). Cách tiếp cận định lượng sử dụng dữ liệu thô chưa được gán nhãn nhằm các mục đích: (a) xác định tần suất xuất hiện tương đối hoặc tuyệt đối của từ, (b) xếp hạng từ dựa vào tần suất xuất hiện, (c) xác định tương đồng về ngữ nghĩa của từ và (d) xác định tần số của chuỗi từ lặp lại (Lemnitzer và Zinsmeister, 2015, tr. 35). Ở cách tiếp cận định tính và định lượng thì "các dữ liệu trích xuất từ khối liệu được phân tích" theo các phạm trù ngữ pháp (không được lấy trực tiếp từ khối liệu) và phân tích dữ liệu là "cơ sở duy nhất cho nghiên cứu" cũng như kết quả phân tích dữ liệu là "nguồn chính dẫn tới các nhận thức về ngôn ngữ" (Lemnitzer và Zinsmeister, 2015, tr. 37).

Tognini-Bonelli (2009) xem hai cách tiếp cận này hoàn toàn đối lập và cách tiếp cận corpus-based đã lỗi thời (dẫn theo Brommer, 2018, tr. 103). Tương tự như vậy là quan điểm của Keibel và cộng sự (2012, tr. 21) khi cho rằng cần phải đi theo "định hướng nghiên cứu việc sử dụng ngôn ngữ thuần túy mà không đưa ra giả thuyết trước", nếu không thì không thể nói tới cách tiếp cận ngôn ngữ học khối liệu. Tuy nhiên quan điểm này bị phê phán ở một số nghiên cứu trong những năm gần đây. Lemnitzer và Zinsmeister (2015, tr. 38) cho rằng quan điểm này là không công bằng. Meißner (2014, tr. 91) lập luận rằng hai cách tiếp cận này cần "đan xen" và "bổ sung lẫn nhau", từ những phạm trù thu được theo corpus-driven có thể xây dựng giả thuyết và kiểm nghiệm giả thuyết theo corpus-based. Theo Brommer (2018, tr. 104), cần phối hợp hai cách tiếp cận này một cách hợp lý "bởi tiềm năng của ngôn ngữ học khối liệu không chỉ đạt được từ những nghiên cứu định lượng, những nghiên cứu này tồn tại một mình thì ít có giá trị khoa học. Thay vào đó dữ liệu thu được từ cách tiếp cận này phải được chọn lựa, phân loại và diễn giải trong các bước phân tích tiếp theo".

3.2. Các phương pháp nghiên cứu

Ngôn ngữ học sử dụng hai phương pháp chính là phương pháp định lượng và phương pháp định tính. Theo Hirschmann (2019, tr. 6-7), các nghiên cứu liên quan đến khối liệu "không bao giờ có tính định lượng thuần túy" và ngược lại, khó có thể tưởng tượng một nghiên cứu "định tính thuần túy". Quá trình phân tích khối liệu bao gồm nhiều bước: tìm kiếm tự động, phân loại, đếm, xác định tần suất, so sánh và phân tích dữ liệu. Để có thể xác định tần suất thì trước đó dữ liệu phải được xử lý và phân loại; sau khi xác định tần suất thì dữ liệu phải được đánh giá và kết quả đánh giá phải được diễn giải. Trong một nghiên cứu định tính thì thông thường nhà nghiên cứu cũng tìm kiếm thông tin về tần suất của các phạm trù được nghiên cứu. Meißner và cộng sự (2016, tr. 309) gọi đó là sự "tích hợp" của nghiên cứu định tính và định lượng. Phần trình bày về hai phương pháp dưới đây chủ yếu dựa vào bài viết của các tác giả này.

Phương pháp định lượng: Mục tiêu của phương pháp phân tích định lượng là xác định tần suất xuất hiện trên cơ sở "đếm số lượng đơn vị", ví dụ đếm tất cả các hình thức xuất hiện của từ (token), các từ khác nhau (type) hoặc các kết hợp từ (collocation). Ví dụ nhóm tác giả Zeldes và cộng sự (2008) (dẫn theo Meißner và cộng sự, 2016, tr. 309) nghiên cứu những cấu trúc gây khó khăn cho người học tiếng Đức bằng cách đếm số lần xuất hiện của tất cả các từ và cấu trúc trong hai khối liệu người học và bản ngữ, từ đó diễn giải những đơn vị từ và cấu trúc ít dùng

trong các văn bản của người học là những lĩnh vực khó khăn. Một ví dụ khác là nghiên cứu tần suất của các từ vựng trong giáo trình dạy tiếng Đức dành cho thanh thiếu niên (Lymperakakis và Sapiridou, 2012, dẫn theo Ahrenholz và Wallner, 2013, tr. 262). Phương pháp phân tích định lượng cho phép "có thể so sánh kết quả trực tiếp với nhau", ví dụ trong nghiên cứu sự phổ biến của từ vay mượn gốc tiếng Anh và tiếng Pháp trong tiếng Đức, O'Halloran (2002) đã xác định tần suất và so sánh chúng: Kết quả cho thấy lượng từ này tăng từ 0,6% (1902) lên 2,0% (1997) trong tổng khối liệu, tương tự thì lượng từ này vào năm 1997 chiếm 4% tổng từ trong ngôn ngữ chuẩn mực, thấp hơn nhiều so với 14% trong ngôn ngữ thời trang (dẫn theo Scherer, 2014, tr. 37).

Các kĩ thuật/công cụ cơ bản trong phương pháp định lượng:

Những ví dụ sau đây được thực hiện dựa trên hai khối liệu bằng phần mềm Antconc: Khối liệu Zeit online[1] gồm 11 bài báo với 5.013 token. Khối liệu truyện cổ tích Grimm gồm 4 truyện với 5.096 token: Aschenputtel (Cô bé lọ lem), Dornröschen (Nàng công chúa ngủ trong rừng), Rotkäppchen (Cô bé quàng khăn đỏ), Rumpelstilzchen (Đồ bỏ xó).

(a) Danh sách từ (wordlist) bao gồm tất cả các dạng thức từ và tần suất của chúng, xem ví dụ ở bảng 2. Sự xuất hiện nhiều nhất có thể được coi là một tiêu chí để phân loại nhóm từ vựng cơ bản và nhóm từ vựng nâng cao.

Bảng 2: Danh sách từ trong khối liệu truyện cổ tích

Rank	Freq	Word
1	295	und
2	188	die
3	105	der
4	101	es
5	99	das

1 Xem tên các bài báo tại danh mục Tài liệu tham khảo.

(b) Từ khóa (keyword) là những từ xuất hiện nhiều hơn trong so sánh với một khối liệu tham chiếu. Phân tích từ khóa có thể sử dụng để nhận diện nhóm từ vựng điển hình của một lĩnh vực sử dụng ngôn ngữ. Trong các từ khóa của khối liệu truyện cổ tích so sánh với khối liệu Zeit online, ta sẽ thấy có khá nhiều từ vựng liên quan đến truyện cổ tích như *Mädchen* (*cô bé*), *Großmutter* (*bà*), *Rotkäpchen* (*cô bé quàng khăn đỏ*), *Aschenputtel* (*cô bé lọ lem*) (xem bảng 3).

(c) Chuỗi từ lặp lại (cluster, n-gram), ví dụ chuỗi 2 từ (bigram), 3 từ (trigram) hay bốn từ (4-gram): Biber và cộng sự (2004) đã dùng kĩ thuật này để so sánh đối chiếu đặc trưng của việc sử dụng ngôn ngữ nói và ngôn ngữ viết trong giảng dạy ở bậc đại học (dẫn theo Meißner và cộng sự, 2016, tr. 309). Trong bảng 4 là một cluster với 4-gram trong khối liệu truyện cổ tích.

Bảng 3: Các từ khóa trong khối liệu truyện cổ tích so sánh với khối liệu Zeit online

Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Keyword Types: 29 **Keyword Tokens:** 1191 **Search Hits:** 0

Rank	Freq	Keyness	Effect	Keyword
9	23	+ 31.56	0.009	mädchen
10	66	+ 30.33	0.0255	sie
11	21	+ 28.81	0.0082	großmutter
12	31	+ 26.46	0.0121	ging
13	19	+ 26.06	0.0074	rotkäppchen
14	18	+ 24.69	0.007	aschenputtel

Bảng 4: Chuỗi từ lặp lại với 4-gram trong khối liệu truyện cổ tích

Concordance | Concordance Plot | File View | Clusters/N-Grams

Total No. of N-Gram Types 4840 **Total No. of N-Gra**

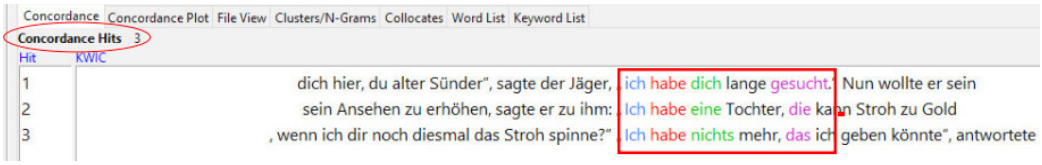
Rank	Freq	Range	N-gram
1	4	1	dass ich dich besser
2	4	1	der könig und die
3	4	1	großmutter was hast du
4	4	1	könig und die königin
5	4	1	was hast du für

(d) Tỷ lệ type và token (type-token ratio TTR) là một đơn vị đo dùng để mô tả biến thể từ vựng hoặc sự đa dạng từ vựng. Tỷ lệ càng tiệm cận 1 thì mức đa dạng càng lớn.

Công cụ này được dùng để đánh giá độ khó của văn bản hoặc mô tả sự phong phú trong cách dùng từ.

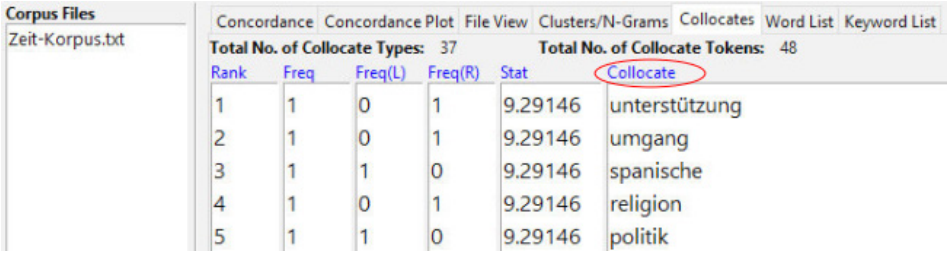
(e) Chỉ mục (concordance) là một kĩ thuật nhằm nghiên cứu các đơn vị từ vựng cần phân tích hoặc nghiên cứu từ khóa trong ngữ cảnh. Tất cả các ngữ cảnh xuất hiện của từ, cụm từ cần phân tích được liệt kê, cho phép xác định các mô hình cấu trúc có chứa từ cần phân tích (xem ví dụ ở bảng 5).

Bảng 5: Cụm từ *ich habe* (*tôi có/tôi đã*) và 4 từ đứng cạnh bên phải (tính từ *ich*)



(f) Kết hợp từ (collocation): Với công cụ này, ta có thể tìm những từ cùng xuất hiện trong ngữ cảnh với từ cần phân tích, xem ví dụ ở bảng 6.

Bảng 6: Kết hợp từ của từ *Regierung* trong khối liệu Zeit online



Phương pháp định tính: Trọng tâm của phân tích định tính là "nghiên cứu sự phức hợp của các hiện tượng ngôn ngữ, nhận biết các qui luật và mô hình mẫu trong sử dụng ngôn ngữ, so sánh chúng với các dữ liệu khác, đồng thời thiết lập các phạm trù xử lý dữ liệu và ứng dụng chúng ở các nghiên cứu tiếp theo" (Meißner và cộng sự, 2016, tr. 312). Mục tiêu của phương pháp phân tích định tính là xác định, phân loại, phân tích và diễn giải những hiện tượng ngôn ngữ hiển thị trong dữ liệu. Ví dụ: Cũng nghiên cứu ảnh hưởng của từ vay mượn gốc tiếng Anh trong tiếng Đức, Schanke (2001) sử dụng phương pháp định tính với mục tiêu xác định sự xuất hiện của các từ gốc tiếng Anh trong khối liệu, tìm ra các từ đó, phân loại theo từ loại và sắp xếp chúng theo các chủ điểm nhất định (trường từ vựng) (dẫn theo Scherer, 2014, tr. 37). Rost-Roth (2006) sử dụng khối liệu để xác định dạng thức và tần suất xuất hiện, phân tích chức năng và phân loại các câu hỏi (dẫn theo Ahrenholz và Wallner, 2013, tr. 262). Khi nghiên cứu lỗi giao thoa văn hóa, nguyên nhân gây lỗi hoặc những khó khăn tiềm ẩn thì việc so sánh với các cấu trúc ở ngôn ngữ thứ nhất thông qua khối liệu song song hoặc khối liệu so sánh (đa ngữ) là cần thiết (Meißner và cộng sự, 2016, tr. 315).

4. Ứng dụng vào lĩnh vực nghiên cứu và giảng dạy tiếng Đức

Khối liệu được ứng dụng vào nhiều lĩnh vực đa dạng, từ ngôn ngữ học (như ngữ pháp, từ vựng học, phương ngữ địa lý, phương ngữ xã hội), ngôn ngữ học lịch sử, từ điển học đến thụ đắc ngôn ngữ và giảng dạy ngoại ngữ, dịch thuật và ngôn ngữ học máy tính (Ahrenholz

và Wallner, 2013, tr. 263-265; Scherer, 2014, tr. 10-15; Lemnitzer và Zinsmeister, 2015, tr. 157-105; Hirschmann, 2019, tr. 7-15). Bài viết này giới hạn ở ứng dụng trong lĩnh vực Tiếng Đức như một ngoại ngữ.

4.1. Câu hỏi và chủ đề nghiên cứu

Có thể nói những vấn đề và câu hỏi nghiên cứu được bàn thảo kĩ lưỡng nhất trong bài viết của Fandrych và Tschirner (2007), trong đó tập trung vào các bình diện ngôn ngữ và các yếu tố đầu vào của việc học tiếng Đức.

Các bình diện ngôn ngữ: Ở bình diện ngữ âm, ngôn ngữ học khối liệu có thể giúp giải quyết những câu hỏi về chuẩn mực phát âm, các biến thể và phương ngữ của người học và tác động của chúng tới quá trình thụ đắc hệ thống âm trong ngôn ngữ đích là tiếng Đức; các vấn đề về tần suất của âm vị và biến thể âm vị, về nuốt âm khi phát âm nguyên âm và phụ âm, về đồng hóa. Ở bình diện hình thái-cú pháp, có thể kể đến hàng loạt các vấn đề nghiên cứu đa dạng như tần suất của các hiện tượng ngữ pháp, quan hệ giữa từ vựng và ngữ pháp, đối chiếu tiếng mẹ đẻ và tiếng Đức, trong đó luôn chú trọng tính đặc thù trong các thể loại văn bản và sự khác biệt trong quá trình tiếp nhận và sản sinh ngôn ngữ. Câu hỏi nghiên cứu cũng có thể liên quan đến cách thức để giúp người học tiếp cận tốt hơn với ngữ pháp tiếng Đức hoặc những phạm trù gây khó khăn cho người học (Hirschmann, 2019, tr. 12-13). Đối với người Việt thì đó là thời của động từ, vị trí của động từ, biến cách của danh từ và tính từ. Ở bình diện văn bản, chủ đề của các nghiên cứu đối chiếu có thể là những qui ước đối với các thể loại văn bản, những tương đồng và dị biệt về văn hóa, xã hội, thể chế, truyền thống khoa học, phân tích cấu trúc vi mô và vĩ mô của văn bản, các phương tiện liên kết văn bản cũng như nghiên cứu về ngữ dụng và giao văn hóa.

Các yếu tố đầu vào bao gồm tần suất, sự nổi bật (salience), sự phức hợp và ngữ cảnh. Fandrych và Tschirner (2007, tr. 200) một mặt phân biệt tần suất của ngôn ngữ nói và ngôn ngữ viết, mặt khác là tần suất của token, type và hình vị. Những tần suất này có thể tác động vào quá trình học tiếng Đức nhưng đồng thời lại có thể xác định dễ dàng nhờ các phương pháp của ngôn ngữ học khối liệu. Liên quan đến sự nổi bật, có thể nghiên cứu việc nhấn mạnh hoặc cảm nhận âm thanh trong ngôn ngữ tự nhiên vận hành như thế nào, những khái niệm ngôn ngữ được thể hiện bằng những phương tiện nào và những phương tiện này khác nhau như thế nào ở tính nổi bật (Fandrych và Tschirner, 2007, tr. 201). Ngôn ngữ học khối liệu cũng có thể góp phần nghiên cứu tính phức hợp về ngữ âm (tần suất của các tập hợp phụ âm), về ngữ pháp (ví dụ như sự biến hình theo giống, số và cách trong danh ngữ gồm quán từ + tính từ + danh từ), về ngữ nghĩa (ví dụ như các dạng thay thế cho các cấu trúc ngữ nghĩa phức tạp và quan hệ của chúng với nhau hoặc những loại nghĩa nào xuất hiện chủ đạo trong những loại văn bản nào). Về ngữ cảnh, Fandrych và Tschirner (2007, tr. 202) đề cập đến các khối liệu người học. Phân tích siêu dữ liệu và ngữ liệu có thể giúp trả lời các câu hỏi về lỗi, chuyển di tích cực và chuyển di tiêu cực, nguyên nhân gây ra lỗi hoặc sự lệch chuẩn.

Đối với các *khối liệu trong ngôn ngữ nói* và trên ví dụ GeWiss[1] (một khối liệu so sánh dành cho việc nghiên cứu và giảng dạy ngôn ngữ nói văn phong khoa học), Fandrych và cộng sự (2018, tr. 6-10) nhấn mạnh các trọng tâm và câu hỏi nghiên cứu sau đây: (a) Bản phiên âm có thể được dùng làm cơ sở nghiên cứu cấu trúc, dàn bài của một bài thuyết trình trong xemina, các qui ước đối với phần mở đầu

[1] Truy cập lúc 10:26 ngày 23.7.2020 tại https://gewiss.uni-leipzig.de/index.php?id=home.

và kết thúc của bài thuyết trình, các phương tiện ngôn ngữ được sử dụng trong đó, quá trình thảo luận (đặt câu hỏi và phản hồi), những đặc điểm điển hình của ngôn ngữ thuyết trình như *okay, unsre, äh, wern* (ja okay also unsre gruppe beschäftigt sich mit dem thema leipzig eine stadt im wandel (1.2) äh zunächst wollen wir erst mal kurz was zum theoretischen hintergrund sagen (.) dann wern wir auf die eigenen äh einzelnen themenkomplexe für den hochschulsommerkurs eingehen …) (Fandrych và cộng sự, 2018, tr. 6-7). (b) Nhờ kĩ thuật tìm chỉ mục, ta có thể nghiên cứu "việc sử dụng các dạng thức ngôn ngữ đặc thù" và chức năng của chúng trong các ngữ cảnh nhất định, ví dụ đánh dấu sự do dự (*äh, ähm*), tín hiệu tiếp nhận (*hm*), biến thể trong khẩu ngữ (nuốt phụ âm cuối *is = ist, nich = nicht, jetz = jetzt* hoặc chỉ phát âm âm cuối *n = ein, ne = eine,* ví dụ: … wenn ma mal drüber nachdenkt wo man ne brücke sehen könnte …) (Fandrych và cộng sự, 2018, tr. 7). (c) Các chú giải về từ loại có thể giúp xác định và nghiên cứu chức năng của các từ nhất định, ví dụ dựa vào tiểu khối liệu về báo cáo hội thảo, có thể nhận thấy tiểu từ *ja* có các chức năng là tín hiệu diễn ngôn (discourse marker) (… gebrauchsliteratur sozusagen °h äh (0.5) und (1.5) ja (.) deren grammatische form und poetische struktur …), tiểu từ tình thái, tín hiệu phản hồi (ja (.) genau.), tín hiệu hỏi xác nhận (ich darf anfangen ja (.) meine sehr verehrten damen und herren …) (Fandrych và cộng sự, 2018, tr. 9). (d) Kĩ thuật từ vị hóa (phục hồi thể nguyên dạng của từ) có thể giúp nghiên cứu việc sử dụng động từ tách, tỷ trọng xuất hiện của dạng thức tách và sự phức hợp của khung câu. (e) Các chú giải về ngữ dụng có thể được sử dụng để nghiên cứu sự thể hiện các hành động bằng ngôn từ như thông báo trước (… das wollen wir dann zumindest kurz an (äh) einigen (.) beispielen an einigen literarischen texten zeigen …), tham chiếu (… was wir ja vorhin schon an dem zitat gesehen haben …), trích dẫn (… so hat harald weinrich bereits in in neunznhundertsiebziger jahren hervorgehoben …), tiểu kết (… was ich f hoffentlich gezeigt habe …) (Fandrych và cộng sự, 2018, tr. 10). (f) Ngôn ngữ học khối liệu cũng có thể cung cấp câu trả lời cho việc thể hiện phong cách cá nhân trong ngôn ngữ khoa học. Ví dụ trong tiểu khối liệu báo cáo hội thảo có tới trên một nửa số chú giải (322/631) cho *ich* (*tôi*) (244) và *wir* (*chúng tôi*) (78) (Fandrych và cộng sự, 2018, tr. 10), kết quả này xác nhận đánh giá của Graefen (1997, tr. 202) cho rằng qui định cấm dùng *ich* thường ít được tuân theo khi tác giả trình bày về kế hoạch và phương pháp nghiên cứu, cách thức tiến hành, các quyết định cá nhân hoặc đưa ra các thông báo.

Đối với *khối liệu người học* thì có 2 cách tiếp cận chính (Hirschmann, 2019, tr. 12): (a) Phân tích mức độ lệch chuẩn trong văn bản của người học so với ngôn ngữ chuẩn mực hoặc các qui phạm và (b) phân tích mức độ lệch chuẩn của một nhóm người học trong đối chiếu với một nhóm khác thông qua một khối liệu so sánh. Lüdeling và cộng sự (2008, tr. 71-72) đưa ra hai ví dụ dựa trên khối liệu Falko: (a) Nghiên cứu định lượng so sánh việc sử dụng liên từ *und* (*và*) đứng ở đầu câu cho thấy ngôn ngữ người học trong nhiều trường hợp lệch so với ngôn ngữ chuẩn, tuy không sai ngữ pháp nhưng là lỗi phong cách, lỗi diễn đạt, hơn nữa tần suất sử dụng cao gấp 8 lần so với người bản ngữ, đặc biệt là ở những câu hỏi tu từ (Und wo haben diese berühmte Menschen ihre Ideen entwickelt, wenn nicht in der Universität?). Những nghiên cứu tương tự cho thấy có thể sử dụng các khối liệu người học để nhận dạng và nghiên cứu những cấu trúc được người học ưu tiên sử dụng (overuse) hoặc tránh không sử dụng (underuse). (b) Khi nghiên cứu lỗi chính tả, các tác giả phân thành hai loại: Lỗi chính tả không phụ thuộc ngữ cảnh (misspelling) (Und wenn

man einen anderen Beispiel nemmen wurde, konnten Frauen zu dieser Zeit ihre Sexualität nicht kontrolieren.*)* và lỗi chính tả phụ thuộc vào ngữ cảnh, nghĩa là chỉ xác định được thông qua ngữ cảnh (syntactical orthography error) (…, aber auf dieser mitleidlosen Welt sollen sie ihre Zierlichkeit nicht verlieren, in dem sie alles versuchen zu machen.). Kết quả nghiên cứu khá bất ngờ: Tỷ lệ lỗi của người học là 111/10.000 từ trong khi ở người bản ngữ là 168/10.000 cũng như tỷ lệ lỗi chính tả không phụ thuộc ngữ cảnh chiếm đa số (92/19), ngược với tỷ lệ ở người bản ngữ (67/101). Những kết quả nghiên cứu dựa trên khối liệu người học như vậy có thể được sử dụng để xây dựng từ điển người học chú trọng tới những đặc thù của ngôn ngữ người học hoặc để cải tiến phương pháp và nội dung giảng dạy (ví dụ lỗi chính tả không đơn thuần chỉ là viết sai mà nguyên nhân chính lại là những vấn đề cú pháp).

4.2. Ứng dụng cho các chủ thể tham gia vào quá trình dạy-học

Giáo viên có thể sử dụng khối liệu vào nhiều mục đích khác nhau (Lüdeling và Walter, 2010b, tr. 3-13; Ahrenholz và Wallner, 2013, tr. 263-264):

(a) Biên soạn học liệu: Khối liệu được sử dụng như một nguồn tham chiếu và nguồn ví dụ. Giáo viên có thể tích hợp những ngữ liệu xác thực, đáng tin cậy và phù hợp cho những tình huống sử dụng đặc thù như trong ngôn ngữ chuyên ngành, ngôn ngữ khoa học, ngôn ngữ chính trị hoặc sử dụng ngữ liệu như một sự bổ sung cho giáo trình (Ahrenholz và Wallner, 2013, tr. 263). Giáo viên cũng có thể dùng ngữ liệu và kết quả nghiên cứu để biên soạn bảng khái quát về các cấu trúc ngữ pháp hoặc phiếu bài tập (Lüdeling và Walter, 2010b, tr. 6-7).

(b) Chữa lỗi: Thường thì giáo viên, đặc biệt nếu không phải là người bản ngữ, sẽ gặp khó khăn khi xác định lỗi ngữ dụng và lỗi phong cách. Trong trường hợp này, giáo viên có thể sử dụng khối liệu như một sự trợ giúp thay vì hoàn toàn dựa vào trực giác (Mukherjee 2002, dẫn theo Lüdeling và Walter, 2010b, tr. 6). Lüdeling và Walter (2010b, tr. 6) cho rằng cần cẩn trọng với đề xuất này vì sự xuất hiện của mỗi ví dụ trong khối liệu không chỉ ra được cấu trúc đó đúng ngữ pháp, ngữ dụng hoặc được chấp nhận hay không cũng như khi một cấu trúc không xuất hiện trong khối liệu thì không có nghĩa là nó không được sử dụng. Tuy nhiên trong một chừng mực nhất định, khối liệu cho phép kiểm chứng sự chấp nhận của những cấu trúc nhất định và giúp giáo viên đưa ra các phản hồi phù hợp (Ahrenholz và Wallner, 2013, tr. 264).

(c) Danh sách xếp hạng từ và cấu trúc theo tần suất có thể giúp cho giáo viên đưa ra quyết định về nội dung giảng dạy bởi thông thường, những cấu trúc và kết hợp từ có tần suất cao sẽ có trọng lượng hơn.

(d) Khối liệu cũng cung cấp cho giáo viên những minh họa về các xu hướng phát triển tiếng Đức. Trong ngôn ngữ nói, đó là việc sử dụng từ vay mượn (*adden, liken, downloaden, updaten*) hoặc động từ trong câu có liên từ *weil* (*vì*), *obwohl* (*mặc dù*) đứng ở vị trí thứ hai (thay vì đứng ở cuối câu) (... weil uns ist was aufgefallen äh als wir den film gemacht hatten den wir eben gezeigt haben; ... aber ich habe doch, durch diesen auftritt bei der pressekonferenz, obwohl ich m muss ja ehrlich sagen, ...) (Ahrenholz và Wallner, 2013, tr. 264; Schneider và cộng sự, 2018, tr. 144, 150).

Đối với *người học* ở trình độ nâng cao, nếu được rèn luyện sử dụng các công cụ tìm kiếm, đặc biệt là công cụ tìm chỉ mục và kết hợp từ, thì khối liệu cũng rất có ích cho các hoạt động sau (Lüdeling và Walter, 2010b, tr. 3-13; Ahrenholz và Wallner, 2013, tr. 264-265):

(a) Kiểm chứng các thông tin về ý nghĩa và chức năng của các đơn vị ngôn ngữ được trình bày trong sách ngữ pháp, từ điển, giáo trình. Tìm hiểu những hiện tượng đặc biệt trong sử dụng ngôn ngữ như đồng nghĩa, đa nghĩa, ngữ trị, giới từ đa nghĩa đa chức năng.

(b) Nhận dạng các cụm từ cố định, các kết hợp từ, ý nghĩa và cách dùng của chúng, qua đó có thể cải thiện các kĩ năng sản sinh.

(c) Ý thức được mối quan hệ giữa chuẩn và biến thể, coi biến thể không phải là lỗi mà là đặc thù của ngôn ngữ nói, ví dụ động từ đứng ở vị trí đầu câu (thay vì đứng ở vị trí thứ 2) (ja, kann schon sein dass man sich auch aus dieser ja privaten krise so n bisschen in den beruf flüchten kann, Schneider và cộng sự, 2018, tr. 182).

(d) Tự xây dựng danh mục tần suất hoặc khối liệu của riêng mình, đặc biệt ở lĩnh vực ngôn ngữ chuyên ngành và dịch thuật.

5. Kết luận

Bài viết đã cho thấy một cái nhìn khái quát về ngôn ngữ học khối liệu ở Đức cũng như tiềm năng lớn của việc ứng dụng vào nghiên cứu và giảng dạy tiếng Đức. Tuy nhiên vẫn còn một số vấn đề cần trao đổi và nghiên cứu để phân ngành khoa học này có thể đáp ứng kì vọng của người sử dụng:

Cần có nhiều nghiên cứu, trao đổi hơn nữa về tiềm năng và giới hạn của việc xây dựng, xử lý, phân tích và ý nghĩa khoa học của khối liệu đối với chuyên ngành Tiếng Đức như một ngoại ngữ, đặc biệt là các nghiên cứu thực nghiệm về các vấn đề liên quan đến quá trình thụ đắc tiếng Đức ở người học với nhiều ngôn ngữ mẹ đẻ khác nhau.

Ngoài những nỗ lực tối ưu hóa nghiên cứu định lượng thì nên tập trung vào những dữ liệu định tính và nghiên cứu định tính. Bên cạnh các nghiên cứu tập trung vào cấu trúc bề mặt thì cần chú trọng các cấu trúc tầng sâu như các hiện tượng về ngữ nghĩa và đa nghĩa, ngữ dụng, hành động ngôn từ, văn phong.

Ở Việt Nam và khu vực Đông Nam Á chưa xây dựng được một khối liệu người học tiếng Đức. Hiện nay, Trường Đại học Ngoại ngữ - ĐHQGHN đang xây dựng một khối liệu người học trong lĩnh vực giảng dạy và nghiên cứu ngôn ngữ Đức và đã có một số nghiên cứu ban đầu dựa vào khối liệu này như nghiên cứu lỗi trong sử dụng liên từ, sử dụng câu phức trong bài thi của sinh viên tiếng Đức trình độ B1-B2, nghiên cứu lỗi trong biên dịch trong cặp ngôn ngữ Đức - Việt. Ngoài ra có một số nghiên cứu đang thực hiện về hệ thống kết hợp từ trong văn bản khoa học của học viên cao học, lập luận có tính nhượng bộ trong tiểu luận và luận văn thạc sĩ ngành Ngôn ngữ Đức v.v. Tuy nhiên đây sẽ chỉ là một khối liệu nhỏ. Cần có những dự án lớn để xây dựng khối liệu của người học ở Việt Nam nói chung và có thể mở rộng sang khu vực.

Ngôn ngữ học khối liệu như một phân ngành hoặc phương pháp cũng chưa được đề cập đến trong nội dung các chương trình đào tạo tiếng Đức. Với những tiềm năng đã nêu trong bài, các nhà nghiên cứu tiếng Đức ở Việt Nam cần bắt đầu nghĩ đến việc ứng dụng phân tích khối liệu và sử dụng các kết quả nghiên cứu để cải tiến phương pháp giảng dạy và biên soạn học liệu.

Ở Việt Nam hiện chưa có các khối liệu tiếng Việt điện tử bất chấp việc phát triển có thể nói rất ấn tượng của kỹ thuật máy tính. Đã đến lúc cần có những đầu tư vào các dự án xây dựng khối liệu tiếng Việt lớn, đáp ứng nhu cầu nghiên cứu và giảng dạy tiếng Việt cũng như so sánh đối chiếu với các ngôn ngữ khác. Đồng thời cũng cần xây dựng và phát triển phân ngành ngôn ngữ học khối liệu và đưa vào giảng dạy tại các trường đại học.

Tài liệu tham khảo

Tiếng Việt

Lê Tuyết Nga (2020). Qui trình xây dựng một khối liệu người học tiếng Đức ở Trường Đại học Ngoại ngữ - Đại học Quốc gia Hà Nội. Kỷ yếu Hội thảo khoa học quốc gia *Nghiên cứu và giảng dạy ngoại ngữ, ngôn ngữ và quốc tế học tại Việt Nam*. Hà Nội, ngày 24 tháng 4 năm 2020, tr. 352-366.

Tiếng Đức

Ahrenholz, B. & Wallner, F. (2013). Korpora für Deutsch als Fremdsprache. In Oomen-Welke, I. & Ahrenholz, B. (Hrsg.), *Deutschunterricht in Theorie und Praxis. (DTP). Handbuch zur Didaktik der deutschen Sprache und Kultur in elf Bänden*. *10. Deutsch als Fremdsprache* (S. 261-272). Baltmannsweiler: Schneider Verl. Hohengehren.

Andresen, M. & Zinsmeister, H. (2019). *Korpuslinguistik*. Tübingen: Narr.

Brommer, S. (2018). *Sprachliche Muster*. Eine induktive korpuslinguistische Analyse wissenschaftlicher Texte. Berlin/Boston: de Gruyter.

Duden. (n.d.). Die häufigsten Wörter in deutschsprachigen Texten. In Duden.de dictionary. Available through <https://www.duden.de/sprachwissen/sprachratgeber/Die-haufigsten-Worter-deutschsprachigen-Texten>, Accessed 17/07/2020 11:00

Fandrych, C., Meißner, C. & Wallner, F. (2018). Das Potenzial mündlicher Korpora für die Sprachdidaktik. Das Beispiel GeWiss. *Deutsch als Fremdsprache*, *55*(1), 3-13.

Fandrych, C. & Tschirner, E. (2007). Korpuslinguistik und Deutsch als Fremdsprache. Ein Perspektivenwechsel. *Deutsch als Fremdsprache*, *44*(4), 195-204.

Greafen, G. (1997). *Der wissenschaftliche Artikel. Textart und Textorganisation*. Frankfurt/Main: Lang.

Hirschmann, H. (2019). *Korpuslinguistik. Ein Einführung*. Stuttgart: Metzler.

Jäkel, O. (2010). The Flensburg English Classroom Corpus (FLECC). Sammlung authentischer Unterrichtsgespräche aus dem aktuellen Englischunterricht auf verschiedenen Stufen an Grund-, Haupt-, Real- und Gesamtschulen Norddeutschlands. Flensburg: Flensburg University Press. Available through <https://www.uni-flensburg.de/fileadmin/content/seminare/anglistik/dokumente/projekte/flecc-online-version.pdf>, Accessed 18/07/2020 15:20

Jones, R. & Tschirner, E. (2006). *Frequency dictionary of German: Core vocabulary for learners*. London: Routledge.

Keibel, H., Kupietz, H. & Perkhuhn, R. (2012). *Korpuslinguistik*. Paderborn: Fink.

Klein, W. P. (2013). Datenbanklinguistik. Eine Weiterentwicklung der Korpuslingustik? In Kratochvílová, I. & Wolf, N. R. (2013) (Hrsg.), *Grundlagen einer sprachwissenschaftlichen Quellenkunde* (S. 333-341). Tübingen: Narr.

Kupietz, M. & Schmidt, T. (2018). *Korpuslinguistik*. Berlin/Boston: de Gruyter.

Lemnitzer, L. & Zinsmeister, H. (2015). *Korpuslinguistik: Eine Einführung*. Tübingen: Narr.

Lüdeling, A. (2007). Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik. In Kallmeyer, W. & Zifonun, G. (Hrsg.), *Sprachkopora - Datenmengen und Erkenntnisfortschritt* (S. 28-48). Berlin/New York: de Gruyter.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K. & Walter, M. (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, *45*(2), 67-73.

Lüdeling, A. & Walter, M. (2010a). Korpuslinguistik. In Krumm, H.-J., Fanrych, C., Hufeisen, B. & Riemer, C. (Hrsg.), *Deutsch als Fremd- und Zweitsprache*. *Handbücher zur Sprach- und Kommunikationswissenschaft* (HSK) 35.1 (S. 315-322). Berlin/New York: de Gruyter.

Lüdeling, A. & Walter, M. (2010b). *Korpuslinguistik für Deutsch als Fremdsprache. Sprachvermittlung und Spracherwerbsforschung* (erweiterte Fassung vom HSK-Artikel). Available through <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/anke/pdf/LuedelingWalterDaF.pdf>, Accessed 25/11/2019 10:15

Meißner, C. (2014). Figurative Verben in der allgemeinen Wissenschaftssprache des Deutschen. *Eine Korpusstudie*. Tübingen: Stauffenburg.

Meißner, C., Lange, D. & Fandrych, C. (2016). Korpusanalyse. In Caspari, D., Klippel, F., Legutke, M. & Schramm, K. (Hrsg.), Forschungsmethoden in der Fremdsprachendidaktik. *Ein Handbuch* (S. 306-319). Tübingen: Narr.

Niederhaus, C. (2011). Die Komplexität von Fachtexten verschiedener Berufsfelder - Eine korpuslinguistische Untersuchung des Fachsprachlichkeitsgrades von Lehrbuchtexten der Berufsfelder Körperpflege und Elektrotechnik. In Granato, M., Münk, D. & Weiß, R. (Hrsg.), Migration als Chance: *Ein Beitrag der beruflichen Bildung (Berichte zur beruflichen Bildung)* (S. 209-224). Bonn: wbv Media.

Paschke, P. (2018). Korpora gesprochener Sprache von/für DaF-LernerInnen. Überblick über mutter- und lernersprachliche Korpora im Kontext von Deutsch als Fremdsprache. In Barbara, V. (Hrsg.), Gesprochene (Fremd-) *Sprache als Forschungs- und Lehrgegenstand* (S. 21-51). Trieste: EUT Edizioni Università di Trieste.

Scherer, C. (2014). *Korpuslinguistik*. Heidelberg: Universitätsverlag Winter.

Schneider, J. G., Butterworth, J. & Hahn, N. (2018*). Gesprochener Standard in syntaktischer Perspektive.*

Theoretische Grundlagen - Empirie - didaktische Konsequenzen. Tübingen: Stauffenburg.

Tschirner, E. (2008). Das professionelle Wortschatzminimum im Deutschen als Fremdsprache. *Deutsch als Fremdsprache*, *45*(4), 195-208.

Tschirner, E. (2009). *Korpuslingustik und Fremdsprachenunterricht.* Germanica Wratislaviensia 129. Acta Universitatis Wratislaviensis, No. 3163 (S. 127-142), Wroclaw. Available through <https://www.researchgate.net/publication/308793121_Korpuslinguistik_und_Fremdsprachenunterricht>, Accessed 17/07/2020 14:15.

Các khối liệu được trích dẫn

Leibnitz-Institut für deutsche Sprache. (n.d.). *Forschungs- und Lehrkorpus Gesprochenes Deutsch*. Available through <https://dig-hum.de/forschung/projekt/forschungs-und-lehrkorpus-gesprochenes-deutsch>, Accessed 18/07/2020 14:20

Ludwig-Maximilians-Universität München. (n.d.). *ALC - Alcohol Language Corpus*. Available through <https://www.phonetik.uni-muenchen.de/forschung/abgeschlossene_projekte/alc.html>, Accessed 18/07/2020 15:00

Merlin Corpus. (n.d.). Available through <https://merlin-platform.eu>, Accessed 18/07/2020 16:00

Universität Hamburg. (n.d.). *Linguistische Profilierung einer europäischen Wissenschaftsbildung (EuroWiss)*. Available through <https://www.slm.uni-hamburg.de/forschung/forschungsprojekte/eurowiss.html>, Accessed 18/07/2020 15:07

Universität Humbold zu Berlin. (n.d.). *Die Falko-Familie: einzelne Korpora*. Available through <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/design>, Accessed 18/07/2020 15:58

Universität Leipzig. (n.d.). *GeWiss. Gesprochene Wissenschaftssprache*. Available through <https://gewiss.uni-leipzig.de/index.php?id=home>, Accessed 23/07/2020 10:26

Universität Stuttgart. (n.d.). *Tiger Korpus*. Available through <https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger>, Accessed 18/07/2020 14:00

Khối liệu truyện cổ tích Grimm

Aschenputtel. (n.d.). In Goethe.de. Available through <http://www.goethe.de/lrn/prj/mlg/mad/gri/de9114281.htm>, Accessed 15/12/2019 10:00

Dornröschen. (n.d.). In Goethe.de. Available through <http://www.goethe.de/lrn/prj/mlg/mad/gri/de9114339.htm>, Accessed 15/12/2019 10:15

Rotkäppchen. (n.d.). In Goethe.de. Available through <http://www.goethe.de/lrn/prj/mlg/mad/gri/de9114344.htm>, Accessed 15/12/2019 10:32

Rumpelstilzchen. (n.d.). In Goethe.de. Available through <http://www.goethe.de/lrn/prj/mlg/mad/gri/de9114371.htm>, Accessed 15/12/2019 10:55

Khối liệu Zeit online

Clan-Kriminalität. Bundesweite Razzia gegen mutmaßliche Schleuserbande. (2019, Oktober 24). Zeit online. Available through <https://www.zeit.de/gesellschaft/zeitgeschehen/2019-10/clan-kriminalitaet-razzia-schleuser-bundespolizei>, Accessed 06/11/2019 10:00

EU-Austritt. Briten sollen bei Brexit-Verschiebung EU-Kommissar stellen. (2019, Oktober 24). Zeit online. Available through <https://www.zeit.de/politik/ausland/2019-10/eu-austritt-brexit-verlaengerung-britischer-eu-kommissar>, Accessed 06/11/2019 10:33

Francisco Franco. Darum wird Spaniens früherer Diktator exhumiert. (2019, Oktober 24). Zeit online. Available through <https://www.zeit.de/gesellschaft/zeitgeschehen/2019-10/francisco-franco-diktator-umbettung-madrid-spanien-faq>, Accessed 06/11/2019 10:55

Husmann, W. (2019, Oktober 23). "Star Wars"-Trailer: Auch das heiligste Skript muss einmal ein Ende haben. *Zeit online*. Available through <https://www.zeit.de/kultur/film/2019-10/star-wars-the-rise-of-skywalker-trailer-release>, Accessed 07/11/2019 20:00

Ilham Tohti. Sacharow-Preis geht an uigurischen Regierungskritiker. (2019, Oktober 24). Zeit online. Available through <https://www.zeit.de/gesellschaft/zeitgeschehen/2019-10/ilham-tohti-sacharow-preis-eu-parlament>, Accessed 07/11/2019 21:00

Immobilienmarkt. Mietpreise stagnieren in mehreren Städten. (2019, Oktober 24). Zeit online. Available through <https://www.zeit.de/wirtschaft/2019-10/immobilienmarkt-miete-wohnen-mietpreise-studie>, Accessed 07/11/2019 21:15

Nordsyrien. USA lehnen Beteiligung an SyrienSchutztruppe ab. (2019, Oktober 24). Zeit online. Available through <https://www.zeit.de/politik/ausland/2019-10/nordsyrien-usa-nato-treffen-schutzzone-mark-esper>, Accessed 07/11/2019 21:20

Opioidkrise. Pharmafirma zahlt 700 Millionen Dollar in Opioidprozess. (2019, Oktober 24). Zeit online. Available through <https://www.zeit.de/wirtschaft/2019-10/opioidkrise-reckitt-benckinser-konsumgueterkonzern-usa>, Accessed 07/11/2019 21:30

Rauterberg, H. (2019, Oktober 24). Leonardo da Vinci. Wunder des Geistes. *Zeit online*. Available through <https://www.zeit.de/2019/44/leonardo-da-vinci-ausstellung-louvre-paris/komplettansicht>, Accessed 07/11/2019 21:38

Rubik's Cube. Zauberwürfel ist doch keine geschützte Marke. (2019, Oktober 24). Zeit online. Available

through <https://www.zeit.de/wirtschaft/2019-10/rubik-s-cube-zauberwuerfel-marke-eu-gericht>, Accessed 07/11/2019 21:55

USA. FBI findet NS-Raubkunst in Museum in New York. (2019, Oktober 24). Zeit online. Available through <https://www.zeit.de/kultur/kunst/2019-10/ns-raubkunst-usa-arkell-museum-new-york>, Accessed 07/11/2019 22:15

CORPUS LINGUISTICS – CONCEPT, APPROACHES, METHODS AND APPLICATIONS IN RESEARCH AND TEACHING OF GERMAN AS A FOREIGN LANGUAGE

Le Tuyet Nga

Faculty of German Language and Culture, VNU University of Languages and International Studies, Pham Van Dong, Cau Giay, Hanoi, Vietnam

Abstract: The paper discusses the concept *corpus* (criteria for corpora, classification of the corpora), corpus linguistics as science or as methodology, approaches (corpus-based approach and corpus-driven approach), research methods (quantitative and qualitative) and tools used in corpus linguistics from the perspective of German scientists. One focus of the work lies in the discussion of the relationship between corpus linguistics and German as a foreign language and in the application possibilities of corpus linguistics for research and teaching of the German language.

Keywords: corpus, corpus linguistics, approach, method, German as a foreign language