Tạp chí Khoa học Trường Đại học Cần Thơ
Phần C: Khoa học Xã hội, Nhân văn và Giáo dục
website: sj.ctu.edu.vn

DOI:10.22144/ctu.jvn.2021.130

ỨNG DỤNG CỦA NGÔN NGỮ HỌC KHỐI LIỆU TRONG NGHIÊN CỨU VÀ GIẢNG DẠY NGOẠI NGỮ QUA VÍ DỤ ĐỐI VỚI TIẾNG ĐỨC

Đặng Thị Thu Hiền*
Khoa tiếng Đức, Trường Đại học Hà Nội
**Người chịu trách nhiệm về bài viết: Đặng Thị Thu Hiền (email: hiendtt@hanu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 24/01/2021
Ngày nhận bài sửa: 14/05/2021
Ngày duyệt đăng: 20/08/2021

Title:
Application of corpus linguistics in studying and teaching of foreign languages through examples of German language

Từ khóa:
Chú giải lỗi, chú giải ngôn ngữ, ngôn ngữ học khối liệu, nghiên cứu thụ đắc ngôn ngữ, khối liệu, khối liệu người học

Keywords:
Error annotation, language annotation, corpus linguistics, language acquisition research, corpora, learner corpora

ABSTRACT

The remarkable development of computer science had a strong influence on the methods of linguistic research in the mid-twentieth and early twenty-first centuries. With the building of corpora including electronic documents representing a certain language, linguists can quickly access and search authentic language materials for their research topics on the basis of linguistic corpora with huge capacity. Various research teams have also designed the learner corpora as they regconized the potential of corpora for teaching and learning foreign languages. This article provides an overview of corpus linguistics, learner corpora and its applicability in research and teaching of foreign languages through examples for German.

TÓM TẮT

Sự phát triển vượt bậc của khoa học máy tính có ảnh hưởng mạnh mẽ đến đường hướng nghiên cứu ngôn ngữ học vào giữa những năm cuối thế kỷ XX và đầu thế kỷ XXI. Với việc xây dựng các ngân hàng ngữ liệu bao gồm các văn bản điện tử đại diện cho một ngôn ngữ nhất định (khối liệu), các nhà ngôn ngữ học có thể nhanh chóng tiếp cận và tìm kiếm ngữ liệu thực cho các đề tài nghiên cứu của mình trên nền tảng các khối liệu có dung lượng khổng lồ. Nhìn thấy được tiềm năng của khối liệu đối với việc giảng dạy và nghiên cứu về giảng dạy ngoại ngữ, nhiều nhóm nghiên cứu trên thế giới cũng đã xây dựng "khối liệu người học". Bài viết dưới đây cung cấp một cái nhìn tổng quan về khối liệu, khối liệu người học và khả năng ứng dụng của nó trong nghiên cứu và giảng dạy ngoại ngữ thông qua ví dụ đối với tiếng Đức.

1. GIỚI THIỆU

Việc tìm kiếm, lựa chọn ngữ liệu đóng một vai trò rất quan trọng trong nghiên cứu ngành ngôn ngữ học. Ví dụ, để có thể mô tả sự phát triển ngữ nghĩa của các đại từ nhân xưng thể hiện sự lịch sự "Sie" trong tiếng Đức, nhà nghiên cứu phải tham khảo một số lượng lớn các văn bản thuộc nhiều thể loại văn phong khác nhau ở các thời điểm lịch sử khác nhau để tìm ra các ví dụ xác thực về việc sử dụng những đại từ này trong những văn cảnh khác nhau. Trước khi có sự trợ giúp của máy tính, việc tìm kiếm này rất khó khăn và tốn nhiều thời gian, đồng thời các ví dụ được tìm thấy mang tính ngẫu nhiên cao; khả năng có thể tìm được các ví dụ mang tính đại diện cho tất cả các văn cảnh có sự xuất hiện của đại từ nhân xưng này là rất thấp. Trong những năm nửa cuối thế kỷ XX và đầu thế kỷ XXI, với sự phát triển mạnh mẽ của công nghệ thông tin, các văn bản được số hóa và được tập hợp một cách hệ thống ngày càng

tăng dẫn đến việc tìm kiếm ngữ liệu đã trở nên dễ dàng hơn. Thay vì phải tìm kiếm một cách thủ công qua việc đọc từng văn bản để tìm ra ví dụ, hiện nay, các nhà ngôn ngữ học có thể tìm được trong tích tắc tất cả các câu/văn cảnh có xuất hiện một đơn vị ngôn ngữ cần trong một tập hợp văn bản có dung lượng lên tới vài tỉ đơn vị từ. Khả năng kỳ diệu này là nhờ các khối liệu đã được xây dựng và không ngừng được mở rộng. Cùng với sự ra đời của các khối liệu là sự hình thành là phát triển của một xu hướng mới trong nghiên cứu ngôn ngữ: ngôn ngữ học khối liệu. Trong các nghiên cứu về thụ đắc ngôn ngữ nói chung và thụ đắc ngôn ngữ thứ hai nói riêng, ngôn ngữ học khối liệu cũng đóng một vai trò ngày càng lớn. Những khối liệu người học ra đời đã mở ra nhiều tiềm năng trong nghiên cứu và giảng dạy/học ngoại ngữ. Bài viết dưới đây đề cập đến các khái niệm cơ bản của ngôn ngữ học khối liệu, khối liệu người học, giới thiệu khối liệu ngôn ngữ Đức (COSMAS II), khối liệu người học tiếng Đức (FALKO, MERLIN) lớn nhất hiện nay và chỉ ra tiềm năng ứng dụng của chúng trong nghiên cứu và giảng dạy tiếng Đức.

2. NHỮNG KHÁI NIỆM CƠ BẢN

Khái niệm "*khối liệu*" (*corpus*) chỉ một tập hợp các văn bản hoặc một phần của văn bản điện tử được lựa chọn và sắp xếp theo những tiêu chí ngôn ngữ nhất định (Scherer, 2006). Văn bản ở đây được hiểu không chỉ là những văn bản ở dạng chữ viết như báo chí, sách hướng dẫn nấu ăn, các tác phẩm văn học, thư từ,... mà còn bao gồm những sản phẩm của ngôn ngữ nói như bài thuyết trình, cuộc nói chuyện tư vấn, bài phát biểu, bài giảng của giáo viên,... Cần lưu ý rằng, văn bản trong một corpus – như trong định nghĩa trên đã chỉ rõ – phải là những văn bản đã được số hóa và có thể tìm kiếm được trên máy tính. Khái niệm corpus được học giả Đào Hồng Thu (2007) nhắc đến lần đầu tiên trong tiếng Việt bằng thuật ngữ "*khối liệu*". Từ những đặc điểm của một *corpus* trình bày ở trên, có thể hiểu "*khối liệu*" chính là một ngân hàng ngữ liệu điện tử của ngôn ngữ nói và viết và có thể đại diện cho một ngôn ngữ (ví dụ ngôn ngữ tiếng Việt, ngôn ngữ tiếng Anh, ngôn ngữ Đức) hoặc một phong cách ngôn ngữ nhất định (ngôn ngữ thanh niên, ngôn ngữ báo chí, ngôn ngữ khoa học trong tiếng Việt).

Khái niệm *corpus* lần đầu tiên được sử dụng như một thuật ngữ khoa học vào năm 1961. Sự ra đời của thuật ngữ này gắn liền với việc xây dựng Brown Corpus (Brown University Corpus of Present-Day American English) – ngân hàng ngữ liệu điện tử đầu tiên trên thế giới với một tập hợp văn bản gồm một triệu đơn vị từ tiếng Anh Mỹ đương đại. Ngay từ khi đó, công nghệ máy tính đã đóng một vai trò then chốt vì nhờ có nó mà việc tìm kiếm những từ, ngữ nhất định trong một khối lượng văn bản khổng lồ có thể thực hiện được trong một thời gian ngắn.

Sự ra đời của Brown corpus có thể nói đã đánh dấu một bước phát triển mới của một xu thế nghiên cứu trong ngành ngôn ngữ học gắn với cái tên *Ngôn ngữ học khối liệu (corpus linguistic)*. Xu hướng này đi ngược lại với phương pháp luận của *Ngôn ngữ học tạo sinh* (*generative linguistic*) do Noam Chomsky (1957) đặt nền móng và vào thời kỳ đó đang có ảnh hưởng mạnh mẽ tại khu vực Bắc Mỹ (Mukherjee, 2009). Đối tượng nghiên cứu của Ngôn ngữ học tạo sinh không phải là những hành vi, lời nói cụ thể (Performance) mà là ngữ năng (Competence) bao gồm những kiến thức trừu tượng của người bản ngữ về quy luật ngôn ngữ tiếng mẹ đẻ của mình. Các nhà ngôn ngữ học thuộc trường phái này thường lấy những câu/lời nói do mình hoặc người bản ngữ tự nghĩ ra để phân tích. Điều này liên quan chặt chẽ đến mục đích của Ngôn ngữ học tạo sinh được Nguyễn Thiện Giáp (2012) tóm lược như sau:

"Kết quả của ngôn ngữ học tạo sinh không phải là miêu tả ngôn ngữ cụ thể, nó lấy ngôn ngữ cụ thể làm điểm xuất phát để tìm ra quy luật chung của ngôn ngữ, cuối cùng làm sáng tỏ hệ thống nhận thức của con người, quy luật tư duy và thuộc tính bản chất của con người."

Ngược lại, mục tiêu nghiên cứu của Ngôn ngữ học khối liệu là việc miêu tả ngôn ngữ được sử dụng thực tế trong một cộng đồng ngôn ngữ nhất định trong một điều kiện giao tiếp tự nhiên. Bởi vậy, ngữ liệu thực được tập hợp trong *corpus* đóng một vai trò đặc biệt quan trọng trong các nghiên cứu theo phương pháp Ngôn ngữ học khối liệu.

Trước khi Brown corpus ra đời thì *Ngôn ngữ học khối liệu* cũng đã luôn là một phương pháp nghiên cứu thực nghiệm của ngành ngôn ngữ học từ nhiều thế kỷ nay. Bản chất của Ngôn ngữ học khối liệu là việc nghiên cứu dựa trên những ngữ liệu thực. Điểm khác biệt của Ngôn ngữ học khối liệu hiện đại mà sự khởi đầu của nó gắn liền với việc xây dựng Brown Corpus so với Ngôn ngữ học khối liệu truyền thống chính là khả năng tìm kiếm tự động những đơn vị ngôn ngữ nhất định trên máy tính nhờ vào những thành tựu vượt bậc của công nghệ thông tin và ngôn ngữ học máy tính (Mukhejee, 2009). Ngôn ngữ học khối liệu hiện đại vì thế còn có cái tên "ngôn ngữ học khối liệu máy tính" (computer corpus linguistics). Theo đó, Ngôn ngữ học khối liệu được

định nghĩa là khoa học nghiên cứu các phương pháp xây dựng và sử dụng khối liệu với sự trợ giúp của công nghệ máy tính (Đào Hồng Thu, 2007).

Tóm lại, khối liệu là thuật ngữ cơ bản của Ngôn ngữ học khối liệu. Theo Đào Hồng Thu (2008) và Mukhejee (2009), khối liệu mang những đặc trưng sau:

1. TÍNH XÁC THỰC: Ngữ liệu được thu thập không phải do nhà nghiên cứu tự nghĩ ra mà là những sản phẩm ngôn ngữ do người sử dụng ngôn ngữ sản sinh trong điều kiện giao tiếp tự nhiên.

2. TÍNH ĐẠI DIỆN: Ngữ liệu được thu thập thuộc nhiều thể loại văn bản khác nhau, của nhiều tác giả, được sản sinh ở nhiều thời điểm lịch sử khác nhau, có tỷ lệ cân bằng, đảm bảo tính phổ quát của kết quả nghiên cứu.

3. TÍNH SỐ HÓA: Khối liệu là một tập hợp các văn bản được số hóa.

4. TÍNH CHÚ GIẢI: Chú giải là "phần giải thích các thông tin đặc thù làm rõ nghĩa cho các văn bản trong khối liệu" (Đào Hồng Thu, 2008), bao gồm chú giải ngoài ngôn ngữ/chú giải ngoại ngôn (*Metadata*) và chú giải ngôn ngữ (*Annotation*). Chú giải ngoại ngôn bao gồm các thông tin về tác giả, hoàn cảnh ra đời của văn bản, thể loại văn bản, quá trình thu thập văn bản. Chú giải ngôn ngữ bản chất là kết quả phân tích văn bản về các bình diện ngôn ngữ, bao gồm chú giải cấu trúc văn bản (đoạn, chương,...), chú giải hình thái học (từ loại, các phạm trù ngữ pháp), chú giải cú pháp (loại câu, thành phần câu, vị trí thành phần câu,...), chú giải ngữ nghĩa và chú giải dụng học. Trong các loại chú giải ngôn ngữ, chú giải hình thái học được coi là nền tảng cho chú giải cú pháp và chú giải ngữ nghĩa (Đào Hồng Thu, 2008). Ngoài các loại chú giải đã nêu còn có chú giải lỗi (xem mục 4.2).

Những đặc trưng kể trên cũng được coi là những yếu tố cần phải cân nhắc đến khi xây dựng khối liệu.

3. ỨNG DỤNG CỦA NGÔN NGỮ HỌC KHỐI LIỆU TRONG GIẢNG DẠY NGOẠI NGỮ

Với sự phát triển nhanh chóng của Ngôn ngữ học khối liệu trong những thập kỷ qua, nhiều Khối liệu đã được xây dựng và phục vụ một cách hiệu quả các nghiên cứu của ngành ngôn ngữ học. Vai trò của Ngôn ngữ học khối liệu đối với việc nghiên cứu và giảng dạy ngoại ngữ cũng đã được khẳng định. Phần trình bày dưới đây giới thiệu Khối liệu tiếng Đức lớn nhất và một số ứng dụng của khối liệu trong nghiên cứu và giảng dạy tiếng Đức.

3.1. Cosmas II

Khối liệu điện tử lớn nhất của tiếng Đức hiện nay là COSMAS II (**Co**rpus **S**earch, **M**anagement and **A**nalysis **S**ystem) do Viện Ngôn ngữ học Đức (IDS) bắt đầu xây dựng từ giữa thập kỷ 60 của thế kỷ XX và liên tục được mở rộng cho đến ngày nay. COSMAS II là thế hệ tiếp theo của COSMAS I và có dung lượng 46,9 tỉ đơn vị từ, tương đương khoảng 130 triệu trang sách (1 trang sách = 400 đơn vị từ) (https://www1.ids-mannheim.de/kl/projekte/korpora/, truy cập ngày 13.10.2020). Trong COSMAS II chứa đựng văn bản viết thuộc nhiều thể loại khác nhau như báo chí, truyện ngắn, văn bản khoa học, khoa học thường thức của tiếng Đức hiện đại và cận hiện đại, trong đó văn bản báo chí chiếm tỷ lệ tương đối cao. Đối với thể loại báo chí, bên cạnh báo chí xuất bản ở Đức còn có báo chí xuất bản tại các nước nói tiếng Đức khác (Áo, Thụy Sĩ). Với dung lượng lớn và với sự phong phú về thể loại văn bản thuộc nhiều lĩnh vực, COSMAS II được coi là Khối liệu đại diện cho Ngôn ngữ Đức và có thể ví nó như 1 lát cắt của tiếng Đức hiện đại.

3.2. Ứng dụng của Khối liệu đối với việc dạy và học tiếng Đức

Xét từ góc độ người nước ngoài học tiếng Đức, Khối liệu như COSMAS II còn được gọi là Khối liệu người bản ngữ (L1-Corpus) để chỉ một tập hợp các sản phẩm ngôn ngữ của người sử dụng tiếng Đức là người bản ngữ. Với loại hình Khối liệu này, có thể tiến hành phân tích định lượng để xác định tần suất sử dụng của các đơn vị ngôn ngữ (từ, cụm từ, cấu trúc ngữ pháp) hoặc phân tích định tính. Nhờ những ứng dụng này, Khối liệu giúp cho việc miêu tả ngôn ngữ một cách chính xác và có thể được sử dụng một cách hữu hiệu đối với việc nghiên cứu ngôn ngữ và giảng dạy ngoại ngữ (Lüdeling & Walter, 2009, Lemnitzer & Zinsmeister, 2015).

Phân tích định lượng: Những Khối liệu được chú giải cho phép thực hiện việc tính toán tần suất của một đơn vị ngôn ngữ nhất định. Ví dụ trên nền tảng Khối liệu tiếng Đức học thuật (Akademisches Deutsch), Lüdeling & Walter (2009) đã xác định được 9 động từ thường (Vollverben) có tần suất cao nhất trong thể loại văn bản khoa học thuộc các lĩnh vực khác nhau (Y học, Ngôn ngữ học, Nông nghiệp); khi đối chiếu với 9 động từ thường được sử dụng thường xuyên nhất trong các bài phát biểu của quốc hội Đức, các tác giả đã khẳng định có sự khác biệt cơ bản giữa 2 thể loại văn bản và qua đó đã tìm ra được những động từ đặc trưng trong thể loại văn bản khoa học. Xuất phát từ giả thiết cho rằng những đơn vị ngôn ngữ nào có tần suất cao là những đơn vị

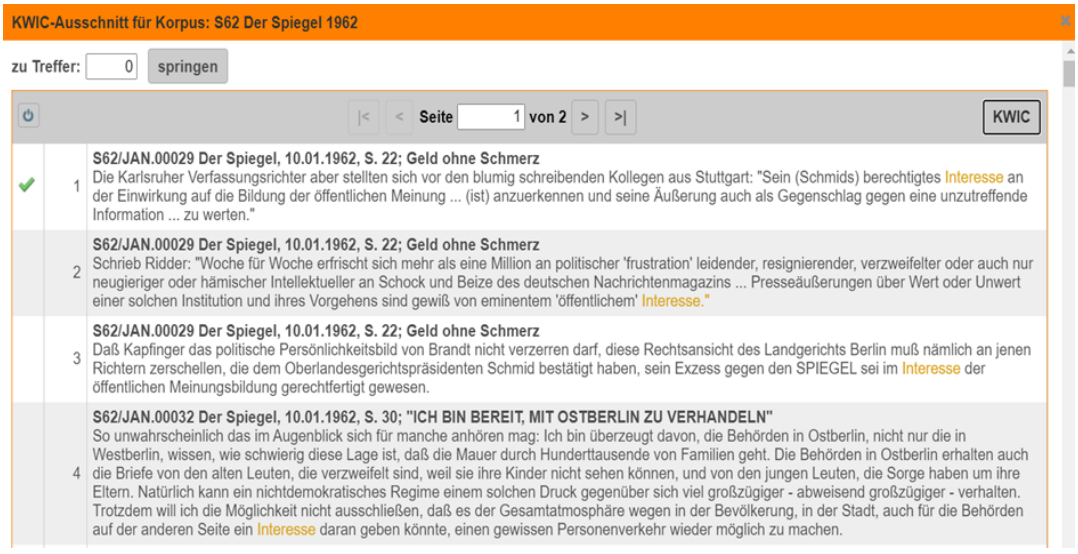
ngôn ngữ phổ thông và do đó cần được dạy cho người nước ngoài, kết quả phân tích về tần suất có thể được dùng làm cơ sở cho việc xây dựng chương trình giảng dạy, biên soạn giáo trình và tài liệu giảng dạy.

Việc tính toán tần suất không chỉ được tiến hành với từng đơn vị từ riêng lẻ mà còn có thể thực hiện với một tập hợp từ. Việc phân tích mức độ thường xuyên của một tập hợp từ (Kollokationsanalyse) cho phép xác định văn cảnh đặc trưng của một đơn vị từ vựng hoặc những đơn vị từ vựng hay xuất hiện cùng nhau. Đây là cơ sở quan trọng cho công tác làm từ điển. Bên cạnh đó, các nhà nghiên cứu/nhà sư phạm học có thể dựa trên những kết quả phân tích như vậy để xác định nội dung giảng dạy hay biên soạn học liệu, đồng thời đưa ra những chỉ dẫn thiết thực cho học viên nước ngoài trong việc ghi nhớ từ cũng như trong việc sử dụng từ đúng văn cảnh (Lüdeling & Walter, 2009; Schmidt, 2010).

Phân tích định tính: Với chức năng tìm kiếm tự động trên nền tảng Khối liệu, kết quả tìm kiếm hiển thị đơn vị ngôn ngữ cần tìm (từ, cụm từ, cấu trúc ngữ pháp) trong cả câu/đoạn văn bản. Tất cả các ngữ liệu của đơn vị ngôn ngữ này hiển thị lần lượt nối tiếp nhau. Cách biểu diễn ngữ liệu như vậy được gọi là *kwic-Konkordanz* (kwic: keyword in context). Hình 1 minh họa giao diện hiển thị kết quả tìm kiếm ngữ liệu cho từ **Interesse** (mối quan tâm) trong Khối liệu tạp chí Tấm gương (der Spiegel) trong COSMAS II (truy cập ngày 17.10.2020):

Hình 1: Giao diện *kwic-Konkordanz* của từ Interesse trên COSMAS II

Đối với mỗi ngữ liệu có từ **Interesse**, Khối liệu có chức năng hiển thị toàn bộ đoạn văn liên quan:



Hình 2. Giao diện thể hiện ngữ cảnh xuất hiện của từ "Interesse" trên COSMAS II

Giáo viên dạy tiếng Đức có thể sử dụng chức năng này của Khối liệu để tìm kiếm nguồn ngữ liệu thực cho nội dung giảng dạy của mình. Ngoài ra, giáo viên dạy tiếng Đức không phải là người bản ngữ có thể coi ngữ liệu trong Khối liệu là một trong những chuẩn mực ngôn ngữ có thể tham khảo làm

cơ sở cho việc chữa bài của học sinh. Trong nhiều trường hợp, có thể học sinh sử dụng cấu trúc ngữ pháp đúng, tuy nhiên không phù hợp về mặt văn phong và văn cảnh. Khi đó, giáo viên có thể kiểm tra sự phù hợp về văn cảnh qua việc nghiên cứu những ngữ liệu thực nhận được từ việc tìm kiếm trên nền tảng Khối liệu. Bên cạnh đó, các đoạn văn bản thực hiển thị ở dạng *kwic-Konkordanz* cũng có thể được dùng để biên soạn các bài tập điền vào ô trống.

Người học tiếng Đức có thể sử dụng ngữ liệu trong Khối liệu là tài liệu học tập, đặc biệt cho việc tự khám phá quy tắc dùng một hiện tượng ngữ pháp hoặc cách dùng một từ vựng chưa biết nghĩa. Để việc học dựa trên Khối liệu (data driven learning) được hiệu quả, cần phải có những khóa tập huấn cho học sinh về Khối liệu và kỹ năng tìm kiếm ngữ liệu trên nền tảng này.

4. KHỐI LIỆU NGƯỜI HỌC VÀ ỨNG DỤNG TRONG GIẢNG DẠY

4.1. Khái niệm "Khối liệu người học"

Đối với các nghiên cứu trong ngành giảng dạy ngoại ngữ, bên cạnh những Khối liệu mà các văn bản là sản phẩm ngôn ngữ của người bản ngữ (L1-corpus), Khối liệu người học (learner corpus hay L2-corpus) cũng đóng một vai trò đặc biệt quan trọng. Căn cứ vào định nghĩa *Khối liệu*, thuật ngữ "*Khối liệu người học*" được hiểu là tập hợp một cách hệ thống các sản phẩm ngôn ngữ đã được số hóa của người học ngôn ngữ (Nesselhauf 2004, trích dẫn bởi Granger, 2008). Trong định nghĩa này, khái niệm "người học ngôn ngữ" được hiểu là người học một ngôn ngữ không phải là ngôn ngữ thứ nhất hoặc là tiếng mẹ đẻ tại nơi mình đang sinh sống. Ví dụ đối với tiếng Đức, khối liệu người học tiếng Đức là tập hợp các văn bản (nói hoặc viết) bằng tiếng Đức của những người học tiếng Đức là ngoại ngữ hoặc là ngôn ngữ thứ hai. Như vậy, đối tượng thu thập của khối liệu người học tiếng Đức có thể là sản phẩm ngôn ngữ Đức của những người có quốc tịch nước ngoài nhưng sinh ra và lớn lên tại nước sử dụng ngôn ngữ đích là tiếng mẹ đẻ, ví dụ người Thổ Nhĩ Kỳ hay người Việt Nam sống tại Đức, hoặc là của những học viên tiếng Đức tại các cơ sở đào tạo ngoại ngữ ở Thổ Nhĩ Kỳ hoặc ở Việt Nam.

Do việc học ngoại ngữ chịu ảnh hưởng của rất nhiều yếu tố nên để có thể được sử dụng cho mục đích nghiên cứu và đảm bảo *tính chính xác* (*Reliability*) của kết quả nghiên cứu thì việc thu thập văn bản cho khối liệu người học phải được lên kế hoạch kỹ lưỡng trên cơ sở cân nhắc các yếu tố liên quan. Bên cạnh những thông tin chung về người học (giới tính, độ tuổi, tiếng mẹ đẻ,...) thì những thông tin liên quan đến môi trường, điều kiện, lịch sử học ngoại ngữ của người học cũng như về văn bản được thu thập (hoàn cảnh ra đời, dạng bài tập, dạng văn bản, chủ đề,...) là những tiêu chí đặc biệt quan trọng cần lưu ý khi xây dựng kế hoạch thu thập *Khối liệu người học* (Granger, 2008).

Khối liệu người học là một loại hình đặc biệt của khối liệu. Việc xây dựng khối liệu người học mới được khởi xướng từ cuối thập niên 80 của thế kỷ 20 (Mukhejee, 2009). Các nghiên cứu với khối liệu người học có thể coi là một nhánh nghiên cứu còn non trẻ của Ngôn ngữ học khối liệu, tuy nhiên đã nhanh chóng khẳng định được vị thế của mình (Granger, 2008).

4.2. Khối liệu người học tiếng Đức

Trong khi có nhiều Khối liệu người học tiếng Anh với quy mô lớn đã được xây dựng và có thể sử dụng miễn phí phục vụ mục đích nghiên cứu khoa học thì việc xây dựng Khối liệu người học tiếng Đức một cách hệ thống để sử dụng rộng rãi có đi sau một bước.

Khối liệu người học tiếng Đức tầm cỡ nhất hiện nay có tên viết tắt FALKO (**F**ehler**a**nnotiertes **L**erner**ko**rpus), do Trường đại học Humbolt (HU) và Trường đại học Tự do Berlin (FU) phối hợp xây dựng. Falko có thể được sử dụng miễn phí qua mạng Internet cho việc nghiên cứu khoa học (https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko). FALKO là một tập hợp sản phẩm viết của người học tiếng Đức ở trình độ từ bậc trung cấp trở lên và bao gồm 2 phần chính sau :

– Bài luận (FALKO-ESSAY) bao gồm 248 bài luận của học viên tiếng Đức có trình độ tối thiểu là B1 được thu thập ở nước ngoài hoặc qua các khóa học mùa hè dành cho sinh viên nước ngoài tại Trường Đại học Tự do Berlin và Trường Đại học Humbold Berlin. Nội dung bài luận xoay quanh 4 chủ đề được gợi ý trước để học viên lựa chọn.

– Bài tóm tắt (FALKO-SUMMERY) bao gồm 196 bài tóm tắt một văn bản khoa học ngành Ngữ văn Đức (văn học hoặc ngôn ngữ học) của sinh viên nước ngoài đang theo học tại Trường Đại học Tự do Berlin. Những sinh viên này có trình độ tiếng Đức tối thiểu đạt bậc C1.

Tương ứng với mỗi Khối liệu thành phần trên của người học tiếng Đức là ngoại ngữ (L2) còn có một Khối liệu đối sánh (Vergleichkorpus) do học sinh/sinh viên nói tiếng Đức là tiếng mẹ đẻ (L1) được thu thập tại các trường đại học và trung học

phổ thông tại Berlin. Các văn bản thuộc Khối liệu đối sánh này có nội dung tương đương với các văn bản trong Khối liệu người học. Việc xây dựng Khối liệu đối sánh này nhằm phục vụ các nghiên cứu so sánh (xem mục 4.3).

Ngoài hai Khối liệu thành phần chính nêu trên, FALKO còn có một Khối liệu thành phần được thu thập tại Trường Đại học Georgtown, Washington (FALKO-GU) bao gồm 92 bài viết của 28 sinh viên Mỹ đang theo học ngành tiếng Đức tại trường. Điểm đặc biệt của FALKO-GU nằm ở chỗ các bài luận của mỗi sinh viên được thu thập trong 3 năm học liên tiếp. Đây là một Khối liệu cắt dọc (Longtudinalcorpus) và phục vụ việc nghiên cứu về sự phát triển năng lực tiếng Đức của cùng một sinh viên ở những giai đoạn/bậc học khác nhau trong quá trình học ngôn ngữ này.

Để có thể tìm kiếm một cách tự động những cấu trúc nhất định phục vụ việc nghiên cứu, ngữ liệu của Falko đã được chú giải. Hệ thống chú giải của Falko là một hệ thống đa cấp; bên cạnh các chú giải từ vựng (Lemmata), từ loại, sự phân đoạn thành phần câu còn có chú giải lỗi. Do phân tích lỗi là một mảng nghiên cứu lớn của chuyên ngành Phương pháp giảng dạy tiếng Đức là ngoại ngữ và việc phân tích lỗi chỉ có thể thực hiện trên cơ sở phân tích sản phẩm ngôn ngữ của người học nên việc chú giải lỗi có thể coi là một loại chú giải đặc biệt quan trọng; là một đặc thù của Khối liệu người học. Chú giải lỗi được thực hiện bằng việc đưa ra một phương án đúng trong ngôn ngữ đích tại những đơn vị ngôn ngữ trong văn bản người học có xuất hiện "lỗi". Một từ/ cụm từ/ câu được coi là lỗi nếu có biểu hiện lệch chuẩn. Bởi vậy, việc đưa ra phương án chuẩn trong ngôn ngữ đích là một bước quan trọng trong việc nhận dạng và phân loại lỗi.

Khối liệu người học tiếng Đức lớn thứ 2 được biết đến tới nay là MERLIN với 1.023 bài thi kỹ năng Viết các trình độ từ A1 đến C1 theo chuẩn chung Châu Âu của thí sinh dự thi kỳ thi năng lực tiếng Đức TELC trên toàn thế giới. Các dạng bài thi Viết bao gồm các thể loại văn bản cá nhân như bưu thiếp, thư điện tử (từ A1 đến B1), văn bản hành chính như thư đề nghị/xin việc/khiếu nại (B2) và văn bản nghị luận/bài báo/báo cáo (C1). Tương tự như FALKO, các văn bản trong MERLIN cũng được chú giải lỗi đa cấp; bên cạnh các chú giải ngôn ngữ (hình vị, cú pháp, từ vựng, chính tả) còn có các chú giải dụng học như tính mạch lạc, văn phong,...

MERLIN là một nền tảng trực tuyến do Liên minh Châu Âu tài trợ và có thể tự do truy cập (https://merlin-platform.eu/C_mcorpus.php#anchor 3). Mục tiêu chính của MERLIN là cung cấp ngân hàng ngữ liệu thực của người học phục vụ việc nghiên cứu đối sánh sự phù hợp của Khung tham chiếu chung Châu Âu về ngôn ngữ (GERS) các trình độ từ A1 đến C1. Ngoài ra, nền tảng này còn có thể được khai thác cho việc xây dựng Chương trình đào tạo, thiết kế tài liệu giảng dạy hay cho việc tự học của học viên học tiếng Đức ở các trình độ cao.

4.3. Ứng dụng của Khối liệu người học trong nghiên cứu thụ đắc ngôn ngữ

Để có thể nghiên cứu về quá trình thụ đắc ngôn ngữ, người nghiên cứu cần phân tích sản phẩm ngôn ngữ đích của người học. Sản phẩm ngôn ngữ này có thể là những sản phẩm người học phải hoàn thành trong quá trình học ngoại ngữ (ví dụ một email, một bình luận ngắn đối với kỹ năng Viết, một bài thuyết trình hoặc một cuộc trò chuyện đối với kỹ năng Nói). Những sản phẩm ngôn ngữ này được tập hợp một cách hệ thống với đầy đủ thông tin về cá nhân cũng như lịch sử học ngoại ngữ của người học trong Khối liệu người học. Bởi vậy, Khối liệu người học có thể coi là một nguồn ngữ liệu quan trọng đối với các nghiên cứu thụ đắc ngôn ngữ thứ hai (Fandrych & Tschirne, 2007).

Nghiên cứu về quá trình thụ đắc ngoại ngữ bao gồm hai mảng nghiên cứu lớn: phân tích lỗi và phân tích đối chiếu.

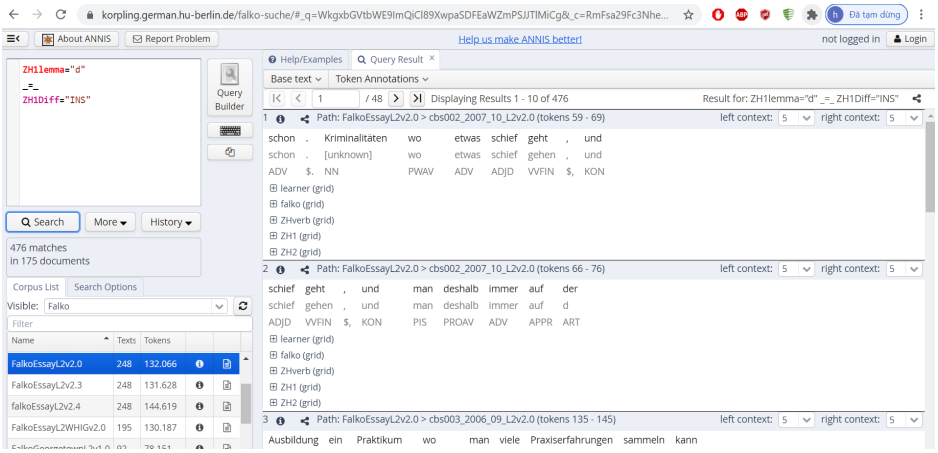
Phân tích lỗi: Việc chú giải lỗi của Khối liệu người học giúp việc tìm kiếm lỗi được thực hiện nhanh chóng, tránh việc nhà nghiên cứu phải phân tích lỗi một cách thủ công.Tùy thuộc bình diện ngôn ngữ được chú giải (chú giải từ vựng, từ loại, ngữ pháp, văn phong) mà khả năng tìm kiếm lỗi ở mỗi Khối liệu người học là khác nhau. Dưới đây là một số ứng dụng của Khối liệu FALKO trong phân tích lỗi.

Là một Khối liệu người học được chú giải từ vựng, từ loại và ngữ pháp nên bên cạnh việc xác định được số lần xuất hiện của một hình thái từ/cụm từ nào đó, FALKO còn cho phép tính toán tự động tần suất của một lỗi về từ vựng hoặc lỗi ngữ pháp trong những bài luận của người học được thu thập trong Khối liệu này qua phần mềm ANNIS:

Ví dụ1:

Câu hỏi: Người học tiếng Đức có thường xuyên mắc lỗi dùng thiếu quán từ trước danh từ không?

Kết quả: Với việc tạo lệnh tìm kiếm tương ứng trong 248 bài luận của người học có trong FalkoEssayL2, ANNIS đã tìm ra được tần xuất mắc lỗi này là 476 lần trong 175 bài luận (https://korpling.german.hu-berlin.de/falko-suche/ (Truy cập ngày 18.10.2020).



Hình 3. Kết quả truy cập trên FALKO về lỗi dùng quán từ

Ví dụ 2:

Câu hỏi: Người học có thường xuyên mắc lỗi đặt sai vị trí của động từ nguyên thể không?

Kết quả: Với việc tạo lệnh tìm kiếm tương ứng trong 248 bài luận của người học có trong FalkoEssayL2, ANNIS đã không tìm thấy lỗi này.

Ví dụ 3:

Câu hỏi: Người học thường xuyên mắc lỗi dùng thừa đại từ nhân xưng "es" khi viết câu không?

Kết quả: Lỗi này xuất hiện trong 33 bài luận và có tổng số 38 lần mắc lỗi được xác định.

Việc tìm kiếm lỗi trên ANNIS có thể được thiết lập cho những câu hỏi nghiên cứu sâu hơn, ví dụ mức độ mắc lỗi tùy theo giới tính, kinh nghiệm học ngoại ngữ, thời gian học tiếng Đức v.v.

Nghiên cứu đối chiếu: Khối liệu người học là nguồn ngữ liệu quan trọng cho việc nghiên cứu so sánh. Ở đây, đối tượng so sánh có thể là một vấn đề/khía cạnh nhất định trong việc học tiếng Đức giữa người học thuộc các quốc tịch khác nhau, hoặc giữa người học và người bản ngữ. Để có thể được sử dụng cho mục đích này, cần có khối liệu người học thuộc các quốc tịch khác nhau và những khối liệu này cần phải được xây dựng dựa trên những tiêu chí tương đối giống nhau (corpus design). FALKO với các khối liệu thành phần được thiết kế giống nhau (xem mục 4.2) cho phép triển khai những nghiên cứu như vậy.

Khi so sánh tần xuất sử dụng của trạng từ "dabei" giữa người học tiếng Đức và người Đức, Schmidt (2010) đã tìm kiếm trên các Khối liệu thành phần của FALKO. Kết quả tìm kiếm đã chỉ ra một số điểm khác biệt cơ bản trong việc sử dụng từ loại này giữa người Đức và người nước ngoài học tiếng Đức. Theo đó, người Đức có xu hướng sử dụng trạng từ này nhiều hơn trên 2 lần so với người đang học ngôn ngữ này, từ đó cho phép kết luận về hiện tượng "dùng ít" (underuse) của từ này. Qua một khảo sát so sánh về tần suất sử dụng một số từ loại trong các khối liệu thành phần của Falko (Khối liệu bản ngữ và Khối liệu người học), Lüdeling & Walter (2009) đã nhận thấy rằng đại từ phản thân *sich* xuất hiện rất ít trong Khối liệu người học với học sinh đến từ nhiều quốc gia khác nhau. Những quan sát này gợi mở cho việc đưa ra giả thuyết cho rằng một số từ loại có thể là một hiện tượng khó đối với người học; bởi vậy họ đã áp dụng "chiến lược né tránh" (Vermeidungsstrategie) trong việc sử dụng chúng nhằm hạn chế việc mắc lỗi. Những nghiên cứu về hiện tượng dùng quá ít một đơn vị ngôn ngữ nào đó đã được học là một hướng nghiên cứu quan trọng trong nghiên cứu thụ đắc ngôn ngữ thứ 2 với mục đích tìm ra những hiện tượng ngôn ngữ khó để từ đó đưa ra những liệu pháp sư phạm nhằm hỗ trợ việc học và sử dụng chúng hiệu quả hơn.

5. KẾT LUẬN

Có thể nói, tiềm năng của khối liệu nói chung và khối liệu người học nói riêng đối với việc nghiên cứu ngôn ngữ và việc dạy/học ngoại ngữ là vô cùng lớn; nó giúp nhà nghiên cứu nhanh chóng tiếp cận với một khối lượng lớn ngữ liệu xác thực đã được chú giải và có thể tìm kiếm tự động. Sự ra đời của các khối liệu như khối liệu đại ngôn ngữ Đức COSMAS II hoặc khối liệu người học tiếng Đức FALKO đã tạo ra một nền tảng ngữ liệu đáng tin cậy và có tính đại diện cao phục vụ nghiên cứu về ngôn ngữ Đức cũng như việc giảng dạy/học tập tiếng Đức. Đối với việc nghiên cứu giảng dạy tiếng Đức ở Việt Nam và nghiên cứu thụ đắc ngôn ngữ Đức của học

sinh Việt Nam, việc xây dựng một Khối liệu người học tiếng Đức của học viên Việt Nam là cần thiết. Hiện tại, Dự án này đang được triển khai tại trường Đại học Ngoại ngữ - Đại học Quốc gia Hà Nội và Trường Đại học Hà Nội, hứa hẹn sẽ cung cấp cho các nhà nghiên cứu nguồn ngữ liệu với sản phẩm ngôn ngữ thực đáng tin cậy cho các đề tài nghiên cứu về việc học tiếng Đức của học viên Việt Nam.

TÀI LIỆU THAM KHẢO

Đào Hồng Thu (2007). Ngôn ngữ học khối liệu (Corpus) (Phần 1). *Ngôn ngữ & Đời sống*, *7* (141), 9-13.

Đào Hồng Thu (2008). Ngôn ngữ học khối liệu (Corpus) (Phần 2). *Ngôn ngữ & Đời sống*, *1+2*(147,148), 23-25.

Granger, S. (2008). Learner corpora. In Lüdeling, A., Kytö, M. (Eds.), *Corpus linguistics. An International Handbook* (pp. 259-274). Walter de Gryter.

Fandrych, Ch. & Tschirne, E. (2007). Korpuslinguistik und Deutsch als Fremdsprache. Ein Perspektivenwechsel (Corpus linguistics and German as a foreign language. A change of perspective). *Deutsch als Fremdsprache*, *44*(4), 195-204.

Nguyễn Thiện Giáp. (2012). Ngôn ngữ học tạo sinh của N. Chomsky: Đối tượng và mục đích. *Ngôn ngữ*, *4*, 3-7.

Lemnitzer, L. & Zinsmeister, H. (2015). *Korpuslinguistik. Eine Einführung (Corpus linguistics. An Introduction)*. Narr Francke.

Lüdeling, A. & Walter, M. (2009). *Korpuslinguistik und Deutsch für Deutsch als Fremdsprache. Sprachvermittlung und Spracherwerbsforschung*. (*Corpus linguistics and German as a foreign language. Language teaching and language acquisition research).* https://www.linguistik.huberlin.de

Mukherjee, J. (2009). *Anglistische Korpuslinguistik. Eine Einführung*. (*English corpus linguistics. An introduction).* Erich Schmidt.

Scherer, C. (2006). *Korpuslinguistik*. (*Corpus linguistics*). Universitätsverlag Winter.

Schmidt, K. (2010). Lernerkorpora: Ressourcen für die Deutsch-als-Fremdsprache-Forschung. (Learner corpora: resources for research of German as a foreign language). In: *Tagungsbeiträge XI. Türkischer Internationaler Germanistik-Kongress* (pp. 555-573). Ege Üniver. Matbaasi.