

ĐẶC ĐIỂM HỒ SƠ TỪ VỰNG CỦA NGỮ VỰC BÀI BÁO KHOA HỌC TIẾNG VIỆT NGÀNH NGÔN NGỮ HỌC

VŨ ANH THU*

Abstract: This study aims to describe the lexical profile of Vietnamese linguistics research articles using a corpus-based approach. The corpus consists of 100 research articles selected from Vietnamese linguistics journals, with the objective of identifying lexical items that occur with high frequency and wide dispersion across the texts. The study adopts a quantitative methodology based on frequency and dispersion measures to determine lexical items that are representative of the target register. The findings indicate that the lexical profile of Vietnamese linguistics research articles is characterized by the predominance of abstract, metalinguistic nouns, verbs associated with cognitive processes and academic presentation, as well as the consistent use of illustrative discourse markers. Methodologically, the study demonstrates that corpus-based and quantitative approaches can be effectively applied to the investigation of Vietnamese academic vocabulary, contributing to greater transparency, consistency, and replicability in linguistic research within the Vietnamese context.

Keywords: *lexical profile, research articles, linguistics, corpus-based analysis*

1. Dẫn nhập

Cùng với sự phát triển của công nghệ và các lý thuyết ngôn ngữ, ngôn ngữ học khối liệu (corpus linguistics) đã mở ra những hướng đi mới mang tính quy mô và đại diện hơn trong mô tả và phân tích ngôn ngữ (Biber, 2006). Không còn phụ thuộc vào trực giác chủ quan của người nghiên cứu hay quy trình thao tác thủ công, cách tiếp cận khối liệu cho phép khảo sát các đặc điểm ngôn ngữ lặp lại có hệ thống trên tập ngữ liệu lớn với các quy trình rõ ràng và thống nhất. Chính từ khả năng xử lý ngữ liệu không lờ và nhanh chóng, nhiều mô thức ngôn ngữ với tần suất cao và phân bố ổn định đã được xác định cho nhiều thể loại ngữ vực (Biber, Conrad & Reppen, 1998).

Một trong những hướng nghiên cứu phổ biến sử dụng phương pháp khối liệu là lexical profile (hồ sơ từ vựng - theo cách dịch của người viết). Khái niệm này được hiểu là tập hợp các đơn vị từ vựng xuất hiện thường xuyên và phân bố rộng trên một khối ngữ liệu cụ thể (Nurmukhamedov & Webb, 2019). Hồ sơ từ vựng không hướng đến mục đích xây dựng danh sách từ mang tính chuẩn hoá hay phổ quát của loại văn bản mà hướng tới việc mô tả các lựa chọn từ vựng tiêu biểu gắn với một ngữ vực hay một lĩnh vực cụ thể. Ở đây, từ vựng chính là nhân tố phản ánh các xu hướng lựa chọn ngôn ngữ mang tính xác suất, hình thành thông qua sự lặp lại có hệ thống trong giao tiếp thực tế.

Với ý nghĩa khoa học trong việc mô tả đặc trưng từ vựng, hồ sơ từ vựng và các nghiên cứu về tần suất cũng như phân bố từ vựng đã thu hút được sự quan tâm lớn của giới học thuật quốc tế. Nhiều công trình đã đi sâu miêu tả hồ sơ từ vựng của các bài báo khoa học, bao gồm vai trò của các danh từ khái quát, động từ học thuật hay các cụm diễn ngôn trong việc tổ chức lập luận và trình bày kết quả nghiên cứu. Mặc dù vậy, phần lớn đối tượng nghiên cứu là ngữ liệu tiếng Anh hay các ngôn ngữ có nguồn ngữ liệu phát triển, trong khi các nghiên cứu ở tiếng Việt còn nhiều hạn chế. Tại Việt Nam, nghiên cứu về từ vựng nói chung hay ngôn ngữ của bài báo khoa học nói riêng thường chỉ xoay quanh đặc điểm phong cách, hệ thuật ngữ hay cấu trúc văn bản cùng với số lượng văn bản hạn chế và phương pháp phân tích thủ công.

Xuất phát từ nhu cầu thực tiễn trên, bài báo này hướng đến mô tả hồ sơ từ vựng của một ngữ liệu nhỏ gồm 100 bài báo khoa học tiếng Việt ngành Ngôn ngữ học theo cách tiếp cận khối liệu thông qua phân tích tần suất và độ phân bố từ vựng, nhằm xác định những đơn vị thường xuyên và ổn định của ngữ vực này. Nghiên cứu không đặt mục tiêu xây dựng danh sách từ chuẩn hoá hay khái quát cho toàn bộ ngữ vực bài báo khoa học mà hướng tới việc thiết lập một khung ngữ liệu cơ bản về hồ sơ từ vựng của bài báo ngành Ngôn ngữ học. Trên cơ sở đó, nghiên cứu góp phần cung cấp bằng chứng định lượng ban đầu về đặc điểm từ vựng của diễn ngôn học thuật này, từ đó làm cơ sở cho các nghiên cứu so sánh, xây dựng danh sách từ học thuật và ứng dụng trong giảng dạy và xử lý ngôn ngữ.

* Trường Đại học Phenikaa; Email: thu.vuanh@phenikaa-uni.edu.vn

Nghiên cứu không nhằm xác định sự khác biệt liên ngữ vực mà tập trung vào việc mô tả có hệ thống các đặc điểm về tần suất và độ phân bố của từ vựng trong nội bộ một diễn ngôn chuyên biệt. Trong bối cảnh các nghiên cứu về từ vựng tiếng Việt học thuật còn hạn chế, việc xây dựng một hồ sơ từ vựng dựa trên khối liệu lớn có ý nghĩa như một nguồn dữ liệu nền, phục vụ cho các nghiên cứu tiếp theo về so sánh ngữ vực, phân tích độ đặc trưng (keyness), cũng như các ứng dụng trong giảng dạy và xử lý ngôn ngữ tự nhiên. Việc giới hạn phạm vi ở một ngữ vực đơn lẻ là lựa chọn có chủ đích nhằm đảm bảo độ sâu phân tích và tính kiểm soát của khối liệu, thay vì mở rộng so sánh khi chưa có một hồ sơ nền đủ tin cậy.

2. Hồ sơ từ vựng, ngữ vực và câu hỏi nghiên cứu

Hồ sơ từ vựng (lexical profile) là phương pháp được sử dụng để tóm tắt và trình bày thông tin về cách sử dụng từ vựng một cách có hệ thống. Nó cho phép phân tích toàn diện dữ liệu để khám phá các mô hình nổi bật (Stubbs, 2001) và được sử dụng để đánh giá việc sử dụng từ vựng hiệu quả trong văn bản viết (Nation & Nation, 2001).

Các từ có tần suất cao là một nhóm từ nhỏ nhưng rất quan trọng vì chúng chiếm tỉ lệ lớn các từ được sử dụng trong các văn bản nói và viết, và chúng xuất hiện trong hầu hết mọi loại hình sử dụng ngôn ngữ (Nation & Nation, 2001). Kiến thức về khối từ này rất quan trọng trong việc sử dụng ngôn ngữ và tiếp cận văn bản nói và viết một cách trôi chảy. Ví dụ: 2.000 từ phổ biến nhất cung cấp độ bao phủ khoảng 86% trong Ngữ liệu Quốc gia Anh (BNC) (Nation & nation, 2001).

Tần suất từ là điểm bắt đầu để xây dựng hồ sơ từ vựng vì nắm được các từ phổ biến là cách tiếp cận dễ dàng nhất để hiểu và sử dụng được ngữ vực nhất định. Bên cạnh đó, hồ sơ từ vựng là hệ thống lựa chọn lặp lại các đơn vị ngôn ngữ (Nation & Nation, 2001). Quan điểm này, được chứng minh bởi bằng chứng từ ngữ liệu, cho thấy hồ sơ từ vựng phản ánh tính quy ước cao và tính lặp lại của các lựa chọn từ vựng (Stubbs, 2001).

Phân tích ngữ liệu cho thấy việc lựa chọn từ không phải là ngẫu nhiên trong văn bản. Ngữ vực được xác lập bằng các đặc trưng ngôn ngữ, các đặc trưng này có tính xác suất và vì vậy, tần suất là nền tảng đầu tiên của ngữ vực. Ngữ vực được định nghĩa bởi một cụm các đặc trưng liên kết có xu hướng đồng xuất hiện (co-occur) với xác suất lớn hơn mức ngẫu nhiên (Biber, 1995). Sự đa dạng về ngữ vực thể hiện sự khác biệt có hệ thống trong việc sử dụng tương đối các đặc trưng ngôn ngữ cốt lõi (chẳng hạn như danh từ so với động từ) (Biber, 2006).

Độ phân tán từ vựng là chỉ báo của ngữ vực. Độ phân tán từ vựng đo lường mức độ đồng đều của một lựa chọn ngôn ngữ phân bố trong một ngữ liệu (Gries, 2008). Chỉ số này cho thấy tầm quan trọng của các từ trong ngữ vực được xét đến. Các từ có độ phân tán thấp, chẳng hạn như từ chỉ xuất hiện trong một số ít văn bản, thường được coi là từ chuyên ngành (specialized). Do đó, nếu một từ có độ phân tán thấp, thì việc sử dụng từ đó chỉ đại diện cho một bộ phận nhỏ chứ không đại diện cho toàn bộ lĩnh vực ngôn ngữ.

Việc kết hợp tần suất và độ phân tán là cần thiết để có được đánh giá chính xác hơn về giá trị của một từ trong ngôn ngữ (Nation & Nation, 2001; Gries, 2008). Tần suất thô (raw frequency) nếu đứng một mình có thể gây hiểu lầm vì nó không tính đến độ phân tán của từ đó. Ví dụ, nếu các từ có tần suất bằng nhau, độ phân tán sẽ chỉ ra rằng chúng có vai trò hay giá trị khác nhau và khi đó, các từ được sử dụng không đồng đều sẽ mang tính chuyên biệt hơn. Khi lựa chọn các từ tần suất cao, cả tần suất và phạm vi phân bố đều phải được xem xét để đảm bảo tính đặc trưng và phổ biến. Phạm vi phân bố là một tiêu chí quan trọng để phân biệt giữa từ có tần suất trung bình và từ có tần suất thấp (Nation & Nation, 2001). Sự kết hợp giữa tần suất và độ phân tán cung cấp thước đo về tính ổn định và tầm quan trọng của một từ.

Với những khái niệm và ý nghĩa của chúng trong nghiên cứu đặc trưng từ vựng của ngữ vực, với mục tiêu mô tả hồ sơ từ vựng của bài báo khoa học tiếng Việt ngành Ngôn ngữ học, bài báo đi tìm câu trả lời cho các câu hỏi nghiên cứu sau:

Câu hỏi 1: Những đơn vị từ vựng nào xuất hiện với tần suất cao trong ngữ liệu?

Câu hỏi 2: Những đơn vị từ vựng này phân bố như thế nào trong ngữ liệu?

Câu hỏi 3: Những từ nào thể hiện tính ổn định và có thể xem là thành tố trong hồ sơ từ vựng của ngữ vực này?

3. Phương pháp và quy trình nghiên cứu

3.1. Thiết kế nghiên cứu

Khảo sát sử dụng phương pháp mô tả định lượng theo hướng tiếp cận khối liệu, cho phép xác định các từ vựng xuất hiện phổ biến với phân bố ổn định trong tập ngữ liệu bài báo khoa học ngành Ngôn ngữ. Theo đó, các khuynh hướng lựa chọn từ vựng thuộc đối tượng nghiên cứu được mô tả dựa trên số liệu thu được.

3.2. Ngữ liệu nghiên cứu

Ngữ liệu nghiên cứu gồm 100 bài báo khoa học ngành Ngôn ngữ học tiếng Việt, được công bố trên Tạp chí Ngôn ngữ và đời sống, xuất bản ở Việt Nam từ năm 2019 đến 2022. Tạp chí Ngôn ngữ và Đời sống là ấn phẩm của Hội Ngôn ngữ học Việt Nam, là một trong những tạp chí uy tín trong lĩnh vực Ngôn ngữ học tại Việt Nam. Tạp chí có chất lượng chuyên môn, tập trung vào đa dạng vấn đề liên quan đến ngôn ngữ và ứng dụng của ngôn ngữ trong đời sống. Bên cạnh đó, các bài đăng trên tạp chí trải qua quy trình phản biện nghiêm ngặt với các nhà khoa học đầu ngành nhằm đảm bảo chất lượng bài đăng và uy tín khoa học của tạp chí. Vì vậy, các bài báo được lựa chọn đều đáp ứng các tiêu chí: (i) là bài báo nghiên cứu khoa học hoàn chỉnh; (ii) được viết bằng tiếng Việt; (iii) thuộc lĩnh vực Ngôn ngữ học; (iv) đảm bảo được chất lượng khoa học và uy tín ngành.

Các bài báo sau khi được lựa chọn ngẫu nhiên được chuyển sang định dạng văn bản thuần túy và lưu dưới dạng các file txt. riêng lẻ. Tất cả văn bản sau đó được làm sạch bằng cách loại bỏ các phần tiêu đề, đề mục, tài liệu tham khảo, hình ảnh, bảng biểu để đảm bảo phân tích chỉ tập trung tuyệt đối vào ngôn ngữ được sử dụng trong bài báo. Số lượng từ trung bình của 100 bài báo sau khi đã được làm sạch khá tương đồng nhau, dao động xung quanh mức 7200 từ. Việc xử lý này nhằm đảm bảo tính nhất quán của dữ liệu và tránh ảnh hưởng của các yếu tố phi ngôn ngữ đến kết quả phân tích từ vựng.

3.3. Công cụ phân tích và xử lý dữ liệu

Để tiến hành phân tách từ và gán nhãn từ loại (POS tagging), một chương trình sử dụng ngôn ngữ lập trình Python và thư viện underthesea (Underthesea – Vietnamese NLP Toolkit) được tạo lập. Công cụ này có khả năng thực hiện phân tách từ, tính toán số lượng và tần suất từ, cũng như gán nhãn từ loại và tổng hợp số liệu theo từng bài báo.

3.4. Quy trình phân tích hồ sơ từ vựng

Quy trình phân tích được tiến hành theo các bước sau:

Bước 1: Tính tần suất từ vựng: Toàn bộ ngữ liệu được đưa vào phần mềm phân tích ngữ liệu và phần mềm thực hiện phân tách từ và gán nhãn từ loại. Các hư từ sau đó được loại bỏ, chỉ giữ lại thực từ. Thực từ là nhóm từ mang nội dung và phản ánh thông tin gắn với chủ đề lĩnh vực. Ngược lại, hư từ mặc dù có ý nghĩa về ngữ pháp và diễn ngôn nhưng tần suất xuất hiện cao và phân bố rộng trong tất cả các ngữ vực nên ít có giá trị trong xác định đặc trưng nội dung của ngữ vực. Kết quả đầu ra là danh sách tần suất từ vựng với thứ tự tần suất giảm dần trong ngữ liệu.

Bước 2: Lựa chọn tập từ có tần suất cao: Từ danh sách tần suất thu được phía trên, nghiên cứu trích xuất 200 từ có tần suất cao nhất để phân tích sâu hơn về độ phân bố. Số lượng 200 chỉ là ngưỡng thao tác ban đầu mang tính phương pháp, nhằm đảm bảo cân bằng được việc bao phủ đủ lượng từ vựng phổ biến và khả năng xử lý định lượng chi tiết ở bước sau.

Bước 3: Phân tích độ phân tán: Các từ trong tập 200 từ phổ biến đầu tiên tiếp tục được phân tích về độ phân tán nhằm xác định mức độ xuất hiện của chúng trong các văn bản khác nhau của ngữ liệu. Một từ được xem là có phân bố rộng nếu nó xuất hiện trong phần lớn các bài báo, thay vì chỉ tập trung ở một số ít văn bản. Ở bước này, phần mềm tiếp tục thống kê số lượng bài mà mỗi từ xuất hiện và từ đó cho ra kết quả là tỉ lệ phân bố của từ trên toàn bộ ngữ liệu. Trong nghiên cứu này, ngưỡng bao phủ 60% lượng văn bản được lựa chọn nhằm đảm bảo tính đại diện xuyên văn bản. Chỉ số phân tán này được sử dụng để tránh chọn các từ thường xuyên nhưng tập trung ở một vài văn bản.

Bước 4: Xác định hồ sơ từ vựng: Dựa trên kết quả phân tích tần suất ở bước 2 và độ phân tán ở bước 3, danh sách các từ vừa có tần suất xuất hiện cao vừa có phân bố rộng trong ngữ liệu được hình

thành và được xem là hồ sơ từ vựng của bài báo khoa học ngành Ngôn ngữ học tiếng Việt trong phạm vi dữ liệu nghiên cứu.

4. Kết quả

4.1. Tập hợp từ vựng có tần suất xuất hiện cao

Kết quả phần này nhằm trả lời câu hỏi nghiên cứu đầu tiên về tập hợp từ vựng có tần suất xuất hiện cao trong ngữ vực bài báo khoa học ngành Ngôn ngữ học tiếng Việt. Kết quả, ví dụ bảng thể hiện 20 từ phổ biến nhất, cho thấy các từ xuất hiện nhiều phần lớn là từ gắn với các hoạt động, hành vi và đối tượng nghiên cứu. Các thuật ngữ chuyên ngành như: *tiếng, thuật ngữ, ngôn ngữ, câu giao tiếp, bị động* gắn với kiến thức lĩnh vực chuyên môn. Các từ *tiếng Anh, tiếng Việt, văn hoá, thí sinh, sinh viên* là những đối tượng hay tham thể nghiên cứu thể hiện chủ đề trong lĩnh vực nghiên cứu hẹp hơn. Tập hợp này chỉ gồm hai loại từ là danh từ (DT) và động từ (ĐT), trong đó danh từ chiếm ưu thế vượt trội (Bảng 1 thể hiện số lượng danh từ gấp ba lần số lượng động từ). Danh sách này cho thấy nội dung của ngữ liệu hướng tới trình bày đối tượng nghiên cứu, phương pháp và kết quả khảo sát, phù hợp với đặc điểm về chủ đề và nội dung của bài báo khoa học ngành Ngôn ngữ.

Bảng 1. Tập hợp 20 từ có tần suất cao nhất

Thứ tự	Từ	Từ loại	Tần suất	Thứ tự	Từ	Từ loại	Tần suất
1	<i>tiếng</i>	DT	244419	11	<i>sử dụng</i>	ĐT	35120
2	<i>tiếng Anh</i>	DT	93836	12	<i>câu</i>	DT	30284
3	<i>thuật ngữ</i>	DT	85476	13	<i>lỗi</i>	DT	29585
4	<i>ngôn ngữ</i>	DT	74055	14	<i>diễn xướng</i>	DT	29189
5	<i>tiếng Việt</i>	DT	63909	15	<i>giao tiếp</i>	DT	27794
6	<i>nói</i>	ĐT	50136	16	<i>viết</i>	ĐT	27393
7	<i>sinh viên</i>	DT	44204	17	<i>bị động</i>	DT	24508
8	<i>hành động</i>	DT	40715	18	<i>văn hoá</i>	DT	24325
9	<i>học</i>	ĐT	40715	19	<i>bài</i>	DT	23660
10	<i>thí sinh</i>	DT	36100	20	<i>dịch</i>	ĐT	21829

4.2. Độ phân bố của các từ phổ biến

Câu hỏi nghiên cứu thứ hai khám phá độ phân bố của các từ xuất hiện thường xuyên nhằm tìm ra mức độ ổn định của chúng trên toàn bộ ngữ liệu. Kết quả cho thấy độ phổ biến không phải lúc nào cũng tỉ lệ thuận với phân bố rộng. Bảng 2 thể hiện tỉ lệ của 20 từ có phân bố rộng nhất trong ngữ liệu. Chẳng hạn như các từ *thuật ngữ* hay *hành động* xuất hiện với tần suất cao (Bảng 1) nhưng lại chỉ tập trung chủ yếu trong một số ít bài báo (34 bài có từ *thuật ngữ* và 51 bài có từ *hành động*). Thậm chí, từ *thí sinh* phổ biến ở bậc 12 nhưng lại chỉ nằm gọn trong một bài báo duy nhất trên 100 bài báo. Ngược lại, các động từ *là, có, thấy, làm, nói, sử dụng, viết, thể hiện, dùng...* hay các danh từ *ngôn ngữ, cách, tiếng, yếu tố...* xuất hiện trong đại đa số các bài báo, là những lựa chọn xuyên suốt mang tính ổn định, thể hiện hành động giải thích, định nghĩa hay các yếu tố cơ bản nền tảng của ngữ vực nghiên cứu.

Bảng 2. Tập hợp 20 từ có độ phân bố rộng nhất

Thứ tự	Từ	Độ phân bố	Thứ tự	Từ	Độ phân bố
1	<i>thấy</i>	97%	11	<i>mới</i>	87%
2	<i>làm</i>	97%	12	<i>thể hiện</i>	86%
3	<i>ngôn ngữ</i>	96%	13	<i>học</i>	85%
4	<i>nói</i>	94%	14	<i>cao</i>	82%
5	<i>cách</i>	94%	15	<i>cần</i>	81%
6	<i>sử dụng</i>	93%	16	<i>nghiên cứu</i>	79%
7	<i>viết</i>	93%	17	<i>dùng</i>	79%
8	<i>tạo</i>	92%	18	<i>yếu tố</i>	79%
9	<i>tiếng</i>	91%	19	<i>hiểu</i>	79%
10	<i>bài</i>	89%	20	<i>câu</i>	74%

4.3. Hồ sơ từ vựng của ngữ vực bài báo khoa học ngành Ngôn ngữ

Câu hỏi nghiên cứu thứ ba hướng đến việc xác định hồ sơ từ vựng của ngữ liệu dựa trên kết quả của tần suất và độ phân bố. Trên cơ sở hai chỉ số độ phổ biến và độ ổn định, nghiên cứu thiết lập một tập hợp các từ vừa xuất hiện thường xuyên vừa phổ biến trong phần lớn văn bản, và được xem là mang tính đại diện cho hồ sơ từ vựng của ngữ vực nghiên cứu. Từ danh sách 200 từ có tần suất cao, nghiên cứu lựa chọn danh sách từ phổ biến rộng ở ngưỡng xuất hiện trên 60% số bài báo trong ngữ liệu. Ngưỡng này là ngưỡng đảm bảo được tính ổn định trên mức trung bình, phù hợp kích thước ngữ liệu mà vẫn không gạt bỏ quá nhiều đơn vị mang tính nội dung học thuật quan trọng, đồng thời, giúp hạn chế ảnh hưởng của các từ gắn với từng lĩnh vực quá chuyên biệt hoặc văn bản riêng lẻ. Với tiêu chí trên, hồ sơ từ vựng của 100 bài báo khoa học ngành Ngôn ngữ gồm 82 từ.

Để có cái nhìn sâu hơn về hồ sơ từ vựng của ngữ liệu nghiên cứu, chúng tôi tiến hành tổng hợp và phân tách các từ trong tập hợp này thành các nhóm nhỏ hơn theo ý nghĩa và vai trò để thấy rõ hơn về khuynh hướng lựa chọn từ vựng của bài báo khoa học ngành Ngôn ngữ. Tập hợp này được chia thành hai nhóm lớn là: nhóm danh từ: danh từ chuyên ngành và danh từ chỉ bối cảnh xã hội; nhóm động từ: động từ chỉ quá trình nhận thức nghiên cứu và động từ thao tác, trình bày. Kết quả được thể hiện trong Bảng 3.

Bảng 3. Phân loại nhóm từ thuộc hồ sơ từ vựng

Loại		Từ	Tần suất
Danh từ	Danh từ chuyên ngành	<i>ngôn ngữ, tiếng, tiếng Việt, tiếng Anh, bài, câu, yếu tố, loại, cấu trúc, ý nghĩa, nghĩa, quan hệ, vai trò, ví dụ</i>	593,833
	Danh từ chỉ bối cảnh xã hội	<i>Việt Nam, xã hội, văn hoá, con người, nước</i>	61,445
Động từ	Động từ chỉ quá trình nhận thức nghiên cứu	<i>nghiên cứu, hiểu, thấy, xem, gọi, xuất hiện, dịch</i>	62,341
	Động từ thao tác, trình bày	<i>sử dụng, viết, đưa, tạo, thể hiện</i>	79,394

Nhóm danh từ phổ biến trong ngữ liệu chủ yếu là danh từ thuộc thuật ngữ chuyên ngành, chỉ đối tượng và yếu tố nghiên cứu ngôn ngữ như *ngôn ngữ, câu, yếu tố, cấu trúc, ý nghĩa...* Đây là những khái niệm quen thuộc của ngành, thể hiện rất rõ xu hướng về chủ đề nghiên cứu của ngành. Ngoài ra, các từ chỉ ngôn ngữ cụ thể như *tiếng Anh* và *tiếng Việt* cũng thể hiện rõ phạm vi nghiên cứu tập trung vào hai ngôn ngữ chính là tiếng Anh và tiếng Việt. Bên cạnh các danh từ chuyên môn, số lượng danh từ chỉ bối cảnh xã hội và văn hoá như *xã hội, Việt Nam, văn hoá* cũng có phổ xuất hiện rộng, phản ánh đặc điểm nghiên cứu ngôn ngữ gắn với mối liên hệ với ngữ cảnh văn hoá và xã hội cụ thể.

Nhóm thứ hai trong hồ sơ từ vựng là nhóm động từ với hai tiểu loại là động từ chỉ quá trình nhận thức nghiên cứu và thao tác, trình bày. Tiểu loại đầu tiên, ví dụ như *nghiên cứu, hiểu, thấy, xem, gọi, xuất hiện, dịch* phản ánh các quá trình nhận thức và tư duy, một hoạt động đặc trưng của khoa học. Thông qua các hoạt động này, các tác giả thể hiện các hành vi quan sát, phân loại, diễn giải trong nghiên cứu ngôn ngữ. Tiểu loại thứ hai gồm các động từ *sử dụng, viết, đưa, tạo, thể hiện* gắn với các hoạt động thể hiện quan hệ, định nghĩa, trình bày phương pháp và diễn giải kết quả, là những yêu cầu quan trọng và mục đích cuối cùng của bài báo khoa học.

Một cách khái quát, hồ sơ từ vựng của bài báo khoa học ngành Ngôn ngữ có số lượng các danh từ chuyên ngành chiếm số lượng áp đảo. Ở đây, đặc trưng về tính học thuật được thể hiện rất rõ và phản ánh thông qua một hồ sơ từ vựng tập trung vào các quá trình định nghĩa, phân loại, diễn giải các hiện tượng, khái niệm ngôn ngữ trong liên hệ với bối cảnh văn hoá - xã hội.

5. Thảo luận

Kết quả phân tích cho thấy các đơn vị từ vựng có tần suất cao trong ngữ liệu chủ yếu tập trung vào hai nhóm chính là danh từ và động từ, trong đó danh từ chiếm ưu thế vượt trội. Điều này phản ánh đặc trưng nội dung của bài báo khoa học ngành Ngôn ngữ học, nơi trọng tâm được đặt vào việc xác định, mô tả và hệ thống hoá các đối tượng nghiên cứu. Sự xuất hiện dày đặc của danh từ cho thấy xu hướng danh hoá, một đặc điểm phổ biến của văn bản khoa học, cho phép làm cô đọng các khái niệm và tăng tính khái quát trong diễn đạt. Bên cạnh đó, các danh từ chỉ đối tượng nghiên cứu cụ thể phản ánh phạm vi nghiên cứu tập trung và định hướng ứng dụng của các công trình trong ngữ liệu.

Tuy nhiên, kết quả cũng chỉ ra rằng tần suất cao không đồng nghĩa với mức độ phân bố rộng. Một số thuật ngữ hoặc từ vựng gắn với nội dung chuyên biệt tuy xuất hiện với tần suất lớn nhưng lại chỉ tập trung trong một số ít văn bản, thậm chí có trường hợp chỉ xuất hiện trong một bài báo duy nhất. Điều này cho thấy sự tồn tại của các tiểu chủ đề trong nội bộ ngành, nơi từ vựng mang tính chuyên biệt cao gắn với từng hướng nghiên cứu cụ thể. Ngược lại, các đơn vị có độ phân bố rộng trên 60% số văn bản lại chủ yếu là những từ đóng vai trò nền tảng trong diễn ngôn khoa học. Những từ này không chỉ thực hiện chức năng ngữ pháp mà còn đảm nhiệm các thao tác diễn ngôn quan trọng như định nghĩa, phân loại, mô tả và diễn giải, những hoạt động cốt lõi trong quá trình kiến tạo tri thức khoa học.

Hồ sơ từ vựng của ngữ liệu cho thấy sự kết hợp giữa hai nhóm chính: (i) danh từ chuyên ngành và danh từ chỉ đối tượng nghiên cứu, và (ii) động từ thể hiện quá trình nhận thức và trình bày khoa học. Sự phân bố của hai tiểu loại này cho thấy bài báo khoa học không chỉ là nơi trình bày kết quả mà còn là không gian thể hiện quá trình tư duy và lập luận của người nghiên cứu. Ngoài ra, sự hiện diện tương đối phổ biến của các danh từ liên quan đến bối cảnh xã hội và văn hoá cho thấy nghiên cứu ngôn ngữ trong ngữ liệu không tách rời khỏi các điều kiện ngữ cảnh cụ thể. Điều này phù hợp với đặc điểm của nhiều hướng nghiên cứu ngôn ngữ học hiện nay, vốn nhấn mạnh mối quan hệ giữa ngôn ngữ và các yếu tố xã hội - văn hoá. Nhìn chung, có thể thấy hồ sơ từ vựng của bài báo khoa học ngành Ngôn ngữ học tiếng Việt được đặc trưng bởi mật độ cao của các danh từ chuyên ngành và các động từ thể hiện quá trình nhận thức và trình bày. Cấu trúc này phản ánh rõ tính học thuật của diễn ngôn khoa học, nơi trọng tâm không chỉ nằm ở việc mô tả hiện tượng mà còn ở việc định nghĩa, phân loại và diễn giải các khái niệm trong một hệ thống tri thức có tổ chức.

Ở góc độ ứng dụng, những kết quả này có thể cung cấp cơ sở ban đầu cho việc xây dựng danh sách từ vựng học thuật tiếng Việt, hỗ trợ giảng dạy viết học thuật cũng như đóng góp dữ liệu cho các nghiên cứu xử lý ngôn ngữ tự nhiên. Tuy nhiên, cần lưu ý rằng do phạm vi ngữ liệu còn giới hạn trong một lĩnh vực và kích thước mẫu chưa lớn, các kết luận này mang tính mô tả ban đầu và cần được kiểm chứng thêm trên các tập ngữ liệu đa dạng hơn.

6. Kết luận

Nghiên cứu này đã cố gắng mô tả hồ sơ từ vựng của một khối ngữ liệu nhỏ gồm 100 bài báo khoa học tiếng Việt ngành Ngôn ngữ dựa trên phân tích tần suất và độ phổ biến từ vựng, áp dụng cách tiếp cận khối liệu và công cụ máy tính. Kết quả cho thấy hồ sơ từ vựng của ngữ liệu này có đặc trưng với ưu thế của các danh từ chuyên ngành và các động từ thể hiện quá trình nhận thức và thao tác nghiên cứu học thuật. Về phương pháp luận, nghiên cứu thử nghiệm này cho thấy hiệu quả của cách tiếp cận khối liệu và tự động, có thể thay thế cho phương pháp thủ công truyền thống, và hoàn toàn có thể cung cấp một bức tranh sơ khai về từ vựng của ngữ liệu cụ thể ngay cả khi chưa có danh sách từ chuẩn hoá hay ngữ liệu đối chứng. Nghiên cứu này không nhằm thay thế các phân tích đối chiếu liên ngữ vực, mà đóng vai trò như một bước khởi đầu trong việc xây dựng cơ sở dữ liệu từ vựng cho ngữ liệu bài báo khoa học tiếng Việt. Trên nền tảng này, các nghiên cứu tiếp theo có thể triển khai các phương pháp như phân tích keyness để xác định rõ hơn các đặc trưng mang tính khu biệt. Đây có thể là bước đầu để làm cơ sở tham khảo và phát triển cho các nghiên cứu với phương pháp hệ thống hơn và ngữ liệu lớn hơn trong tương lai đối với tiếng Việt.

TÀI LIỆU THAM KHẢO

1. Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.
2. Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam/Philadelphia: John Benjamins.
3. Biber, D., Conrad, S., & Reppen, R. (1996). *Corpus-based investigations of language use*. Annual review of applied linguistics, 16, 115-136.
4. Gries, S. T. (2008). *Dispersions and adjusted frequencies in corpora*. International journal of corpus linguistics, 13(4), 403-437.
5. Nation, I. S., & Nation, I. S. P. (2001). *Learning vocabulary in another language* (Vol. 10, pp. 126-132). Cambridge: Cambridge university press.
6. Nurmukhamedov, U., & Webb, S. (2019). *Lexical coverage and profiling*. Language Teaching, 52(2), 188-200.
7. Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. John Wiley & Sons.