NGÔN NGỮ HỌC NGỮ LIỆU - HÀNH TRÌNH TỪ TRUYỀN THỐNG ĐẾN HIỆN ĐẠI

CORPUS LINGUISTICS - A JOURNEY FROM TRADITION TO MODERNITY

Nguyễn Thị Thúy *

Ngày tòa soạn nhận được bài báo: 03/01/2022
Ngày nhận kết quả phản biện đánh giá: 04/07/2022
Ngày bài báo được duyệt đăng: 27/07/2022

Tóm tắt: *Ngôn ngữ học ngữ liệu là một ngành khoa học khá mới mẻ đối với các nhà nghiên cứu ngôn ngữ - văn hóa Việt Nam. Nhằm mục đích xác định các giai đoạn phát triển của Ngôn ngữ học ngữ liệu, bài viết hệ thống, tổng hợp và mô tả vắn tắt các nghiên cứu nổi bật ở Việt Nam và trên thế giới về lịch sử hình thành và phát triển của ngành khoa học này. Dựa vào vai trò của máy tính điện tử trong các nghiên cứu của Ngôn ngữ học ngữ liệu, bài viết chia lịch sử Ngôn ngữ hoc ngữ liệu thành hai giai đoạn: Ngôn ngữ học ngữ liệu truyền thống và Ngôn ngữ học ngữ liệu hiện đại. Kết quả xác định và phân đoạn lịch sử Ngôn ngữ học ngữ liệu của bài viết là một tham khảo cho các nhà nghiên cứu khi tìm hiểu về Ngôn ngữ học ngữ liệu nói chung và lịch sử ngành khoa học này nói riêng.*

Từ khóa: *Ngôn ngữ học ngữ liệu, kho ngữ liệu, nguồn ngữ liệu, lịch sử, lịch sử ngôn ngữ học ngữ liệu.*

Abstract*: Corpus linguistics is a relatively new science for Vietnamese researchers. In order to identify the development stages of corpus linguistics, the article synthesizes and briefly describes the outstanding researches in Vietnam and the world on the history of formation and development of corpus linguistics. Based on the role of electronic computers in corpus linguistics studies, the article divides the history of linguistics into two periods: traditional corpus linguistics and modern corpus linguistics. The results of identifying and segmenting the history of corpus linguistics in the article are a reference for researchers when studying about corpus linguistics in general and the history of this science in particular.*

Keywords: *corpus linguistics; corpus, corpus source, history, history of corpus linguistics.*

I. Đặt vấn đề

Trong quá trình khảo cứu các tài liệu để thực hiện việc xây dựng một kho ngữ liệu tiếng Việt về kinh tế phục vụ cho việc giảng dạy và nghiên cứu, chúng tôi đã nhận thấy có một sự bất nhất trong việc xác định lịch sử hình thành và phát triển của ngành Ngôn ngữ học ngữ liệu. Rất nhiều nhà nghiên cứu hiện nay cho rằng, Ngôn ngữ học ngữ liệu là một ngành khoa học trẻ và sự hình thành của nó không thể thiếu vai trò của khoa học máy tính nói riêng, công nghệ thông tin nói chung. Nhiều nhà khoa học khác lại cho rằng, Ngôn ngữ học ngữ liệu đã ra đời từ rất lâu

* Trường Đại học Kinh tế quốc dân

khi chưa có công nghệ máy tính và điều đó có nghĩa là Ngôn ngữ học ngữ liệu không hẳn là một khoa học trẻ, hình thành và tồn tại gắn liền với máy tính điện tử.

Trong bài viết này, trên cơ sở thảo luận về nền tảng của Ngôn ngữ học ngữ liệu và xem xét các nghiên cứu về Ngôn ngữ học ngữ liệu, chúng tôi đưa ra quan điểm về lịch sử hình thành và phát triển của ngành khoa học này. Bài viết đóng góp vào những tài liệu về Ngôn ngữ học ngữ liệu còn đang khá hạn chế ở Việt Nam.

II. Cơ sở lý thuyết

2.1. Lý thuyết về kho ngữ liệu

Thuật ngữ *kho ngữ liệu* (corpus, số nhiều là corpora) có nguồn gốc từ tiếng Latin, có nghĩa là body (thân thể). Thuật ngữ *kho ngữ liệu* được ghi nhận xuất hiện từ năm 1961 với sự ra đời của kho ngữ liệu điện tử đầu tiên, ngữ liệu Brown. Tuy nhiên như chúng tôi tìm hiểu, kho ngữ liệu đã được xây dựng và khai thác từ rất lâu trước đó. Các kho ngữ liệu trước kho Brown chủ yếu được thu thập, lưu trữ, xử lí thủ công. Một kho ngữ liệu thường cần phải đảm bảo ba tiêu chí: tính xác thực *(Authenticity)*, tính đại diện *(Representativeness)*, và kích cỡ *(Size)*. Trong các tiêu được coi là quan trọng nhất của kho ngữ liệu rất ít nhà khoa học đề cập đến tính chất điện tử. Chính vì vậy, chúng tôi đưa ra một định nghĩa về kho ngữ liệu như sau: *"Kho ngữ liệu là một tập hợp lớn các mẫu văn bản nói hoặc (và) viết được sử dụng trong thực tế, được lựa chọn một cách có hệ thống và dựa vào các tiêu chí nhất định, được xây dựng theo cách thủ công hoặc điện tử, nhằm phục vụ cho việc nghiên cứu ngôn ngữ và các công việc khác có liên quan".*

2.2. Lý thuyết về Ngôn ngữ học ngữ liệu

Thuật ngữ Ngôn ngữ học ngữ liệu (*Corpus linguistics*) được sử dụng lần đầu bởi Aarts và Van den Heuvel vào năm 1982, nhưng theo Léon (Ramesh Krishnamurthy & Wolfgang Teubert, 2007), phải đến những năm 1990, thuật ngữ này mới được sử dụng rộng rãi với sự gia tăng nhanh chóng của các ấn phẩm và đặc biệt là sự ra đời của tạp chí International Journal of Corpus Linguistics (IJCL).

Nhiều nhà nghiên cứu hiện nay cho rằng, Ngôn ngữ học ngữ liệu là một ngành khoa học trẻ và sự hình thành của nó không thể thiếu vai trò của khoa học máy tính nói riêng, công nghệ thông tin nói chung. Mc Enery (2012) định nghĩa "*Ngôn ngữ học ngữ liệu là khoa học nghiên cứu dữ liệu ngôn ngữ trên quy mô lớn - phân tích bộ sưu tập phong phú các bản phiên âm lời nói hoặc văn bản viết có sự hỗ trợ của máy tính*". Tác giả Đào Hồng Thu cũng có cái nhìn tương tự về vai trò của máy tính điện tử trong sự hình thành nên Ngôn ngữ học ngữ liệu. Tác giả cho rằng "*Ngôn ngữ học khối liệu (thuật ngữ tác giả sử dụng tương đương với thuật ngữ kho ngữ liệu trong bài viết) là giao điểm giữa khoa học ngôn ngữ và khoa học máy tính, được hình thành vào cuối thế kỉ 20 trên cơ sở điện tử kĩ thuật số, là khoa học nghiên cứu xây dựng các khối liệu ngôn ngữ, nghiên cứu các phương pháp xử lí dữ liệu và sử dụng khối liệu*" (Đào Hồng Thu, 2007). Một số nhà nghiên cứu khác không đưa máy tính điện tử vào trong định nghĩa về Ngôn ngữ học ngữ liệu. Chẳng hạn, Sadinha (2004) cho rằng, Ngôn ngữ học ngữ liệu "*tập trung vào thu thập và vận dụng kho ngữ liệu, hoặc là một bộ dữ liệu ngôn ngữ được thu thập cẩn thận, để phục vụ như một nguồn lực nghiên cứu ngôn ngữ hoặc các biến thể ngôn ngữ*" (dẫn theo Carlos, 2019). Nguyễn Thiện Giáp (2016) đưa ra định nghĩa như sau: "Ngôn ngữ học kho ngữ liệu là sự nghiên cứu ngôn ngữ như được biểu lộ trong các mẫu của văn bản thực".

III. Phương pháp nghiên cứu

Phương pháp đầu tiên được sử dụng trong bài viết là phương pháp tổng hợp, hệ thống. Phương pháp này được áp dụng trong việc thu thập, sắp xếp lại các nghiên cứu từ trước đến nay về Ngôn ngữ học ngữ liệu, qua đó đưa ra những nhận định về vai trò của chúng trong lịch sử nghiên cứu Ngôn ngữ học ngữ liệu. Phương pháp so sánh đối chiếu các nghiên cứu cũng được áp dụng nhằm mô tả lại các giai đoạn lịch sử của ngành khoa học này một cách khái quát, thống nhất. Nguồn tham khảo chính của bài viết là các công trình của các nhà nghiên cứu về Ngôn ngữ học ngữ liệu có thời gian xuất bản tương đối gần với thời điểm hiện tại, cụ thể là từ năm 2000 đến nay.

IV. Kết quả và thảo luận

Nếu theo cách nhìn coi khoa học máy tính như thành phần thiết yếu của Ngôn ngữ học ngữ liệu thì lịch sử của khoa học này chỉ có thể được xác định bắt đầu từ những năm giữa của thế kỉ 20, khi kho ngữ liệu điện tử đầu tiên, kho ngữ liệu Brown, ra đời.

Nhưng những quan điểm về các kho ngữ liệu điện tử không thể phủ nhận một thực tế rằng, thành phần trung tâm trong Ngôn ngữ học ngữ liệu chính là bộ sưu tập ngữ liệu, cụ thể hơn là ngữ liệu được sản sinh ra trong bối cảnh sử dụng thực tế và Ngôn ngữ học ngữ liệu là khoa học nghiên cứu về ngôn ngữ thông qua các dữ liệu ngôn ngữ xác thực đó. Trước khi có máy tính, các nhà khoa học ngữ liệu cũng đã có rất nhiều nghiên cứu đi theo hướng tiếp cận này. Xét từ đối tượng đến cách tiếp cận, việc xem xét lịch sử của Ngôn ngữ học ngữ liệu được tính bắt đầu khi ngữ liệu được điện tử hóa là không thỏa đáng và thiếu sót.

Theo khảo sát, chúng tôi chia lịch sử hình thành và phát triển của ngôn ngữ học ngữ liệu thành hai giai đoạn: giai đoạn trước 1960 và giai đoạn từ 1960 đến nay. Giai đoạn trước 1960 có thể tính bắt đầu từ thế kỉ 13, là giai đoạn Ngôn ngữ học ngữ liệu truyền thống với đặc trưng thủ công trong việc thu thập, xử lí ngữ liệu. Giai đoạn từ 1960 đến nay được gọi là Ngôn ngữ học ngữ liệu hiện đại khi công nghệ máy tính bắt đầu có những can thiệp và sau đó là tham dự vào như một thành phần quan trọng trong tất cả các công việc của Ngôn ngữ học ngữ liệu từ thu thập, lưu trữ, xử lí, truy cập, đến khai thác ngữ liệu.

4.1. Ngôn ngữ học ngữ liệu truyền thống

Tính về mặt thời gian, Ngôn ngữ học ngữ liệu truyền thống có một tiến trình dài hơi so với Ngôn ngữ học ngữ liệu hiện đại. Tuy nhiên, do tính chất thủ công trong tất cả các công đoạn nên những nghiên cứu của Ngôn ngữ học ngữ liệu truyền thống còn hạn chế về số lượng cũng như phạm vi áp dụng. Các công việc chủ yếu là xây dựng các danh sách các từ (chỉ mục từ) có kèm theo ngữ cảnh sử dụng phục vụ chủ yếu cho tra cứu từ trong Kinh thánh, nghiên cứu trong văn học, so sánh ngôn ngữ, và biên soạn từ điển.

Từ thế kỉ 13, những thao tác mà các nhà Ngôn ngữ học ngữ liệu ngày nay vẫn làm như là chú thích để tìm kiếm từ hoặc cụm từ gắn với ngữ cảnh trong một khối văn bản lớn đã được thực hiện bởi một nhóm các học giả Kinh thánh. Các học giả này đã lập danh sách các từ trong cuốn Kinh thánh Cơ đốc theo thứ tự aphabet cùng với các trích dẫn (chỉ mục – concordance) nơi mà các từ đó được sử dụng nhằm phục vụ cho việc tra cứu đồng thời chứng minh rằng Kinh thánh là một thông điệp thần thánh thống nhất chứ không phải là sự kết hợp của một loạt các văn bản khác nhau. Những công việc tương tự cũng được thực hiện bởi

linh mục Cardinal Hugo, vào năm 1230, với sự hỗ trợ của một đội ngũ 500 tu sĩ, đã xây dựng chỉ mục từ về luân lí của cuốn Kinh thánh Vulgate (phiên bản Latinh thế kỷ thứ năm của kinh thánh). Như vậy có thể thấy, những công việc vẫn được cho là quen thuộc đối với Ngôn ngữ học ngữ liệu hiện đại sau này có nguồn gốc từ các công việc tỉ mỉ của các học giả kinh thánh suốt từ thế kỉ 13 đến giữa thế kỉ 20.

Sau giai đoạn này, tiếp tục có nhiều nghiên cứu khác về ngữ liệu mở rộng ra trên những lĩnh vực khác. Cuốn từ điển tiếng Anh của Samuel Johnson, *A Dictionary of the English Language*, bắt đầu thực hiện từ 1746 xuất bản năm 1755 là kết quả của tám năm tác giả làm việc với kho ngữ liệu trên giấy, một kho ngữ liệu với vô số ghi chép tỉ mỉ bằng tay về các ví dụ ngôn ngữ trong sử dụng từ giai đoạn 1560 đến 1660. Đây có lẽ là ví dụ nổi tiếng nhất về kho ngữ liệu thủ công lưu trữ trên giấy với hơn ba triệu từ vựng được ghi lại cùng ngữ cảnh của nó. Cuốn từ điển của Johnson được đánh giá là tốt nhất thời đại của nó và một trong những đổi mới chính của từ điển là bao gồm các trích dẫn nổi tiếng từ văn học và các nguồn khác để chứng minh ý nghĩa và cách sử dụng từ trong ngữ cảnh. Từ điển tiếng Anh Oxford nổi tiếng sau này đã sao chép khoảng 1.700 định nghĩa của Johnson, đánh dấu chúng đơn giản là 'J.' (Johnson) như một hành động chứng minh sự ảnh hưởng của cách tiếp cận và phương pháp biên soạn từ điển của tác giả. Cùng thời gian với Samuel Johnson cũng có một loạt các nghiên cứu khác. Năm 1787, Becket tạo một danh sách chỉ mục ngữ cảnh của các từ trong các tác phẩm của Shakespeare, cung cấp một nguồn lực quý giá cho các nhà nghiên cứu văn học thời bấy giờ. Năm 1897, J. Kading sử dụng một kho ngữ liệu tiếng Đức gồm 11 triệu từ để tính toán sự phân bố tần số các chữ cái trong từ vựng tiếng Đức. Năm 1907, W. Stern ghi chép lại toàn bộ ngôn từ của trẻ từ lúc bắt đầu bập bẹ đến lúc lớn để nghiên cứu khả năng nhận biết ngôn ngữ của trẻ em. Từ 1909 – 1949, Otto Jespersen nghiên cứu và xuất bản bảy tập sách *A Modern English Grammar on Historical Principles* về ngữ âm, hình thái học, ngữ pháp tiếng Anh không thể không tính đến vai trò của kho ngữ liệu văn học Anh mà ông đã trích dẫn trong đó hàng nghìn ví dụ để minh họa cho các cấu trúc mà ông thảo luận trong các nghiên cứu của mình. Năm 1947, H. Bongers đã khai thác từ ngữ liệu để rút ra danh sách các từ thông dụng nhất để phục vụ cho việc học ngoại ngữ. Năm 1952, V.Fries dựa trên ngữ liệu để nghiên cứu ngữ pháp tiếng Anh theo hướng mô tả (descriptive grammar). Nửa đầu của thế thế kỉ 20, đặc biệt vào những năm 40, 50, nghiên cứu trong ngôn ngữ học theo cách tiếp cận của Ngôn ngữ học khối liệu rất sôi động đặc biệt được khẳng định trong công việc của các nhà ngôn ngữ học cấu trúc trước Chomsky mà tiêu biểu là các nhà ngôn ngữ học thuộc trường phái Cấu trúc luận Mĩ. Họ quan niệm *"Tổng hợp các phát ngôn có thể được phát ra trong một cộng đồng ngôn ngữ là ngôn ngữ của cộng đồng ấy"*, (dẫn theo Nguyễn Thiện Giáp, 2012) nên chủ trương nghiên cứu các hiện tượng ngôn ngữ mà người nghiên cứu có thể quan sát được, đó chính là các ngữ liệu họ thu thập được.

Dù thể hiện được vai trò quan trọng trong nghiên cứu và nhiều lĩnh vực khác với những thành tựu đột phá nhưng sự thật là Ngôn ngữ học ngữ liệu giai đoạn này rõ ràng phải chấp nhận đó là sự giới hạn của sức người trước khối lượng công việc khổng lồ khi xây dựng một kho ngữ liệu và hiệu quả khai thác kho ngữ liệu áp dụng cho nghiên cứu. Chính điều này đã khiến Ngôn ngữ học ngữ liệu chưa được nhìn nhận đúng với vai trò của nó.

4.2. Ngôn ngữ học ngữ liệu hiện đại.

Công nghệ máy tính là đặc điểm khác biệt đầu tiên để nhận diện hai giai đoạn trong lịch sử phát triển của Ngôn ngữ học ngữ liệu. Sử dụng máy tính để lưu trữ ngữ liệu có thể đã xuất hiện từ những năm 1950 với công nghệ máy tính thẻ đục lỗ (Punched card). Tuy nhiên, không phải quốc gia nào cũng đủ phát triển để tạo ra và vận hành cỗ máy công nghệ khổng lồ, đắt đỏ và đòi hỏi người dùng có trình độ cao đó. Năm 1959, R. Quirk cùng với các cộng sự ở Đại học London (University College London) đã thành lập Survey of English Usage (SEU) – một trung tâm nghiên cứu lớn về ngữ liệu ở Châu Âu nhưng những ngữ liệu tiếng Anh sử dụng trong thực tế của SEU ban đầu cũng vẫn được ghi lại vào các cuộn băng hoặc chép lại trên giấy. Mục đích của SEU là cung cấp các nguồn tài liệu để mô tả chính xác ngữ pháp được sử dụng bởi những người bản ngữ. Dự án này vẫn tiếp tục suốt những năm sau đó và đến năm 1975, Đại học Lund, Thụy Điển đã tiếp tục công trình của Quirk như một giai đoạn thứ hai của dự án khi tiếp tục thu thập, lưu trữ, xử lí ngữ liệu và máy tính hóa khối ngữ liệu đã được thu thập từ trước và phải đến lúc này nó mới được biết đến rộng rãi và có giá trị hơn với tên London-Lund Corpus (LLC).

Cùng khoảng thời gian với SEU, ở Mỹ, máy tính dùng thẻ nhớ đục lỗ phát triển hơn nhiều và đây là lí do dù SEU rất nổi tiếng với hoạt động thu thập và điều tra ngữ liệu ở Châu Âu, nhưng kho ngữ liệu điện tử đầu tiên lại ra đời ở Mỹ, đó chính là kho ngữ liệu Brown (Brown Corpus - Brown University Standard Corpus of Present-Day American English, 1961). Brown không chỉ là kho ngữ liệu điện tử đầu tiên mà nó còn là kho ngữ liệu được tổ chức một cách khoa học với 500 mẫu văn bản tiếng Anh – Mỹ, mỗi mẫu khoảng 2000 từ, tổng kho hơn một triệu từ, đại diện cho nhiều thể loại khác nhau. Brown Corpus được kế thừa về mặt cấu trúc, cách thức xây dựng bởi hàng loạt các kho ngữ liệu khác sau này: kho ngữ liệu Lancaster – Oslo/Bergen về tiếng Anh - Anh (LOB,1970s), kho Freiburg – Lancaster-Oslo/Bergen về tiếng Anh – Anh (FLOB, 1990s); kho Freiburg-Brown về tiếng Anh – Mỹ (FROWN, 1990s); kho Crown về tiếng Anh – Mỹ (2009). Hệ thống các kho ngữ liệu trên nền tảng Brown Corpus được gọi là *Brown family corpora*.

Khoảng thời gian hơn mười năm từ cuối những năm 1950 cho đến đầu thập niên 70 đánh dấu mốc Ngôn ngữ học ngữ liệu chuyển mình sang giai đoạn hiện đại, nhưng cũng thời kì này Ngôn ngữ học ngữ liệu phải nhận sự phê phán kịch liệt của N. Chomsky, một nhà ngôn ngữ học đại diện tiêu biểu cho chủ nghĩa lí luận trong ngôn ngữ và tham vọng xây dựng một chủ nghĩa câu trúc mới bằng lí thuyết ngữ pháp tạo sinh. Cuốn sách về những tư tưởng của ông *Syntactic Structures* (1957) gây ảnh hưởng rộng khắp và lôi cuốn các nhà ngôn ngữ học thời bấy giờ. Ở một phía khác, có thể thấy, từ những nghiên cứu đầu tiên đến giai đoạn hiện tại (và hẳn nhiên cả sau này), Ngôn ngữ học ngữ liệu cơ bản theo tinh thần của chủ nghĩa thực nghiệm (một phong trào trong triết học Phương Tây đối lập với chủ nghĩa lí luận – cũng là phong trào ảnh hưởng lớn đến nghiên cứu ngôn ngữ nửa đầu thế kỉ 20). Chủ nghĩa thực nghiệm cho rằng chỉ những phát biểu có thể kiểm chứng được thông qua quan sát trực tiếp hoặc bằng chứng logic mới có ý nghĩa, ngược lại với chủ nghĩa lí luận cho rằng tri thức có được là từ trực giác, lí luận của nhà nghiên cứu. Mặc dù chủ nghĩa thực nghiệm vẫn nhận được sự ủng hộ trong nghiên cứu khoa học nói chung và ngôn ngữ học nói riêng, nhưng những tư tưởng của Chomsky đã đưa Ngôn ngữ

học quay lại với chủ nghĩa duy lí bằng hàng loạt các giả thuyết và lập luận đầy khác biệt nhưng hấp dẫn về tính bẩm sinh ngôn ngữ của trẻ em, về ngữ pháp phổ quát chung cho loài người, về cấu trúc sâu. Lí thuyết ngôn ngữ của Chomsky đã chiếm vị trí chủ đạo trong ngôn ngữ học trong suốt ba thập kỉ 60, 70, 80 và gây nhiều sức ép đối với Ngôn ngữ học ngữ liệu trong thời kì vận động chuyển mình, khiến ngành khoa học này đã chững lại trong khoảng hơn hai thập niên.

Thực sự thì từ thập niên 70, những phát triển vượt bậc của công nghệ thông tin với sự ra đời của những máy tính có tốc độ xử lí, khả năng lưu trữ mạnh hơn ngàn lần so với những năm 50-60, sự ra đời của Internet kết nối đã bắt đầu kích hoạt lại nhu cầu xây dựng và nghiên cứu về ngữ liệu như nghiên cứu của Sinclair và các cộng sự tại đại học Birmingham trong những năm 60. Tuy nhiên, phải sang thập niên 80, khi công nghệ thông tin đã có thêm một thập niên để cái tiến khả năng lưu trữ, chức năng phần mềm; đặc biệt, khi máy tính cá nhân ra đời (máy tính cá nhân đầu tiên là Acorn của IBM, 1981), Ngôn ngữ học ngữ liệu đã phục hồi mạnh mẽ, ngoạn mục với hàng loạt các kho ngữ liệu lớn nhỏ và các nghiên cứu có liên quan. Có thể nói, máy tính điện tử đã giải phóng cho Ngôn ngữ học ngữ liệu, tạo ra cuộc cách mạng cho ngành khoa học, nó khiến những công việc trước đây hoặc tưởng là không thể thực hiện được hoặc mất quá nhiều công sức và thời gian để thực hiện thì nay trở nên dễ dàng và hiệu quả hơn rất nhiều. Giờ đây, các nhà ngôn ngữ học có thể thu thập hàng nghìn trang dữ liệu trong một thời gian ngắn, tìm kiếm một phần nhỏ bất kì của ngôn ngữ qua kho dữ liệu đó trong vài giây, soạn những bộ từ điển khổng lồ dựa trên cách sử dụng từ ngữ trong thực tế, nghiên cứu và đưa ra những kết luận chính xác về ngôn ngữ dựa trên những dữ liệu đã được thiết lập và xử lí tỉ mỉ, vận dụng kết quả khảo sát kho ngữ liệu vào giảng dạy, dịch thuật và cho chính việc cái tiến năng lực của máy tính chẳng hạn như khả năng xử lí ngôn ngữ tự nhiên (Natral Language Processing- NLP). Máy tính điện tử như một lần nữa "khai sinh" Ngôn ngữ học ngữ liệu nên dễ hiểu vì sao các nhà nghiên cứu hiện nay thường coi công nghệ của máy tính như một phần không thể thiếu của Ngôn ngữ học ngữ liệu nói chung, của kho ngữ liệu nói riêng.

Sự cải tiến của máy tính cũng đã tạo ra những cuộc chạy đua trong xây dựng ngữ liệu. Dự án kho ngữ liệu Cobuild (Đại học Birmingham) do J. Sinclair khởi xướng và chủ trì vào khoảng những năm đầu thập kỉ 80, chứa 10 triệu từ, nhưng một số kho ngữ liệu xuất hiện gần như ngay sau đó đã được đẩy lên hàng trăm triệu từ. Đến những năm giữa thập niên đầu tiên của thế kỉ 21, kho ngữ liệu Cambridge International (Cambridge University Press) đã nâng con số này lên một tỉ từ. Năm 2016, kho ngữ liệu News on the Web (Now – Đại học Brigham Young) dữ liệu của nó đã lên tới 13,3 tỉ từ và nó đang được cập nhật hằng ngày. Không khó để có thể hình dung về những kho ngữ liệu có thể chứa toàn bộ dữ liệu của các trang web trong tương lai không xa.

Bên cạnh các kho ngữ liệu khổng lồ như trên, các nhà nghiên cứu cũng xây dựng những kho ngữ liệu kích thước nhỏ nhưng được tổ chức, xử lí cẩn thận, khai thác các phần mềm chức năng nhằm đáp ứng hiệu quả mục đích nghiên cứu của nhà khoa học. Chẳng hạn như các kho ngữ liệu trong hệ thống kho ngữ liệu Brown. Xuất phát từ mô hình của kho Brown đầu tiên (1961), các thành viên sau đó của Brown family cũng đều chứa trong nó khoảng một triệu từ, với mỗi mẫu ngữ liệu là 2000 từ,

nhưng là các biến thể khác nhau của tiếng Anh như tiếng Anh – Anh, Anh – Mỹ ở từng thời kì khác nhau. Dữ liệu của kho được phân thành các thể loại và được gán nhãn từ loại, từ ghép, từ viết tắt, từ mượn… Ngữ liệu PTB (Pennsylvania Tree Bank) cũng được coi là một kho ngữ liệu "vàng" của tiếng Anh được xây dựng với hơn bảy triệu từ được gán nhãn từ loại, gán nhãn cú pháp. Đây là kho ngữ liệu mà hầu hết các chương trình gán nhãn từ loại hay cú pháp sử dụng để huấn luyện máy tính.

Cùng với sự ra đời các kho ngữ liệu điện tử hiện đại là những kết quả ứng dụng trên mọi lĩnh vực. Một số lĩnh vực tiêu biểu có thể kể đến như từ điển học, ngữ pháp học, nghiên cứu ngôn ngữ, giảng dạy ngôn ngữ.

Đầu tiên phải kể đến từ điển học, lĩnh vực mà từ giai đoạn Ngôn ngữ học ngữ liệu truyền thống đã đạt được những thành tựu to lớn. Một loạt các từ điển được biên soạn dựa trên ngữ liệu như *Collins COBUILD English Dictionary* (dựa trên kho Bank of English Corpus), *Cambridge International Dictionary of English* (dựa trên kho Cambridge International Corpus và kho Cambridge Learners' Corpus); *Longman Dictionary* (dựa trên kho British National Corpus).

Trên địa hạt ngữ pháp, không thể không nhắc đến *A Comprehensive Grammar of the English Language* (Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, Jan Svartvik, 1985). Đây là cuốn sách ngữ pháp tiếng Anh được soạn trên nguồn ngữ liệu của ba kho ngữ liệu sớm nhất The Survey of English Usage (1959), The Brown Corpus (US English) – 1960s, The Lancaster-Oslo-Bergen Corpus (UK English) - 1970s,. Nó được coi là cuốn sách ngữ pháp mô tả (descriptive grammar - ngữ pháp dựa trên quan sát cách dùng trên thực tế rồi khái quát hóa thành quy tắc) tiếng Anh đương đại *"vĩ đại nhất" "kĩ lưỡng và chi tiết nhất""vượt ra khỏi ranh giới quốc gia"* (https://en.wikipedia.org/wiki/A_Comprehensive_Grammar_of_the_English_Language) Nhiều sách ngữ pháp sau Quirk thậm chí còn dựa nhiều hơn vào ngữ liệu, như *Greenbaum's Oxford English Grammar* (1996) được trích dẫn từ International Corpus of English (ICE-GB); *Longman Grammar of Spoken and Written English* (1999) dựa vào Longman Spoken and Written English Corpus.

Nghiên cứu ngôn ngữ có lẽ là lĩnh vực ứng dụng kho ngữ liệu rộng rãi nhất. Tùy vào mục đích nghiên cứu, các nhà khoa học sẽ tự xây dựng kho ngữ liệu riêng hoặc lựa chọn kho ngữ liệu phù hợp. Chẳng hạn, dựa vào hai kho ngữ liệu Brown (Anh – Mỹ) và LOB (Anh – Anh), hai kho ngữ liệu được thiết kế tương đương với nhau về mô hình và thời gian của ngữ liệu, các nhà khoa học đã so sánh đồng đại giữa hai biến thể được sử dụng phổ biến nhất của tiếng Anh là Anh – Anh và Anh – Mỹ. Nhưng để so sánh lịch đại, lựa chọn lại là Brown, LOB trong đối sánh với Frown, hay FLOB, những kho tiếng Anh – Anh, Anh – Mỹ trong khoảng thời gian từ 1980 – 1990. Các nhà nghiên cứu đã phân tích và đưa ra nhiều kết luận giá trị về cách sử dụng từ (Enery và Xiao, 2004, 2005), tần suất từ loại (Mair và cộng sự, 2002), phân biệt ngôn ngữ nói và ngôn ngữ viết (Hudson, 1994; Rayson, 1997, Granger và Rayson 1998, Biber, 1999) các đặc điểm ngữ pháp (Leech và Smith, 2006), phân tích diễn ngôn (Aijmer và Stenstrom¨ 2004; Baker 2006; Biber,1998), ngữ nghĩa học (Ensslin và Johnson, 2006), ngôn ngữ học xã hội (Gabrielatos và cộng sự, 2010) … (tham khảo Mc Enery, 2012).

Giảng dạy ngoại ngữ cũng là lĩnh vực mà Ngôn ngữ học ngữ liệu có những can thiệp tạo ra những đổi mới về tài nguyên cũng như phương pháp. Nguồn tài nguyên

để giảng dạy ngoại ngữ có thể là các kho ngữ liệu đơn ngữ, cũng có thể là các kho ngữ liệu song ngữ, hay các kho ngữ liệu người học (Learner's Corpus). Một số kho ngữ liệu người học nổi tiếng International Corpus of Learner English (ICLE, 1990), kho ngữ liệu Longman Learner …Các kho ngữ liệu này đã cung cấp những thông tin vô giá về lỗi khi học tiếng Anh, các lỗi nào là điển hình, tần suất các lỗi đối với các đối tượng người học khác nhau đã giúp tạo ra các tài liệu để học tiếng Anh (Granger, 2003). Các kho ngữ liệu cũng tác động rất lớn đến các lĩnh vực chuyên biệt hơn như giảng dạy tiếng Anh chuyên ngành (English for Specific Porpose – ESP) (Mohamad-Ali, 2007), thiết kế giáo trình (Mindt 1996; Shortall 2007), đánh giá ngôn ngữ (Alderson, 1996; Taylor và Barker 2008), thực hành giảng dạy trên lớp (Amador-Moreno, 2006), tài liệu tham khảo, hướng dẫn học tập thông qua ngữ liệu (Johns 1994, 1997; Boulton 2009) (theo Mc Enery, 2012).

V. Kết luận

Ngôn ngữ học ngữ liệu mới chỉ được biết đến ở Việt Nam khoảng hai thập kỉ nên các công trình nghiên cứu về ngữ liệu còn hạn chế. Bài viết tổng quan các nghiên cứu nổi bật dựa vào ngữ liệu từ trước đến nay và qua đó xác định các giai đoạn hình thành và phát triển của Ngôn ngữ học ngữ liệu. Bài viết chia lịch sử Ngôn ngữ học ngữ liệu thành hai giai đoạn, giai đoạn truyền thống từ 1960 đổ về trước và giai đoạn hiện đại từ 1960 đến nay dựa trên tính chất điện tử hóa của ngữ liệu. Máy tính đã không bắt đầu cùng với sự ra đời của Ngôn ngữ học ngữ liệu, nhưng chắc chắn rằng nó đã hồi sinh và thúc đẩy mạnh mẽ cho những đóng góp của Ngôn ngữ học ngữ liệu trên tất cả các lĩnh vực nghiên cứu và trở thành một phương pháp nghiên cứu không thể thiếu đối với nhiều ngành khoa học đặc biệt là các khoa học liên quan về ngôn ngữ.

Tài liệu tham khảo:

Tiếng Việt

[1]. Bình, L. T. (2016). *Nghiên cứu xây dựng kho ngữ liệu giáo khoa tiếng Anh chuyên ngành Xã hội học. Luận án tiến sĩ.* Hà Nội: Trường Đại học Khoa học Xã hội và Nhân văn – Đại học Quốc gia Hà Nội.

[2]. Điền, Đ. (2018). *Ngôn ngữ học khối liệu.* Thành phố Hồ Chí Minh: NXB Đại học quốc gia Thành phố HCM.

[3]. Giáp, N. T. (2016). *Từ điển khái niệm Ngôn ngữ học.* Hà Nội: NXB Đại học Quốc gia, Hà Nội.

[4]. Hiển, P. (2006). *Sử dụng kho ngữ liệu trong giảng dạy tiếng Việt.* Hà Nội: Từ điển học & Bách khoa thư, Số 1.

[5]. Phúc, T. H. (2017). *Nghiên cứu điều kiện "IF" biểu hiện chiến lược lịch sự trong diễn ngôn báo chí Anh bằng phương pháp khối liệu.* Kỉ yếu hội thảo khoa học Quốc gia 2017: Nghiên cứu và giảng dạy ngoại ngữ, ngôn ngữ và quốc tế học tại Việt Nam.

[6]. Thu, Đ. H. (2007). *Ngôn ngữ học khối liệu (Corpus).* Hà Nội: Số 7, Tạp chí Ngôn ngữ và đời sống.

Tiếng Anh

[1]. Carlos Assunção, Carla Araújo. (2019). *Entries on the history of corpus linguistic.* Sao Paulo.

[2]. Meyer, C. F. (2002). *English Corpus Linguistics - An Introduction* . New York: Cambridge University Press.

[3]. O'Keeffe, A. (2010). *Historycal perspective: What are coppora and how have they envolved? (The Routledge Handbook of Corpus Linguistics).* London: Routledge.

[4]. Stefanowitsch, A. (2020). *Corpus linguistics A guide to the methodology.* Berlin: Language Science Press.

[5]. Tony McEnery, A. H. (2012). *Corpus Linguistics Method, Theory and Practice.* New York: Cambridge University Press.

Địa chỉ tác giả: Trường Đại học Kinh tế quốc dân

Email: thuyngth@neu.edu.vn