NGÔN NGỮ HỌC KHỐI LIỆU TRONG NỀN KINH TẾ TOÀN CẦU

LINGUISTIQUE DE CORPUS DANS L'ECONOMIE MONDIALE

TS. Đào Hồng Thu[1]

Tóm tắt

Sự ra đời và phát triển mạnh mẽ của công nghệ thông tin đã dẫn tới sự ra đời và phát triển của hàng loạt các lĩnh vực hoạt động khoa học và công nghệ khác, trong đó có các lĩnh vực hoạt động của ngôn ngữ học. Song song với sự phát triển không ngừng của các thế hệ công nghệ máy tính và dịch tự động, trong ngôn ngữ học hình thành xu hướng phát triển mới - ***Ngôn ngữ học khối liệu (Corpus Linguistics)****.*

Thực tế đã chứng minh rằng ngôn ngữ học khối liệu ngày càng đóng vai trò quan trọng trong một nền kinh tế. Bài báo đề cập đến vai trò của ngôn ngữ học khối liệu - một khoa học xuất hiện vào nửa cuối thế kỉ XX vừa qua - như một đóng góp của Khoa học xã hội và nhân văn trong phát triển kinh tế-xã hội và sự phát triển của khoa học về khối liệu ngôn ngữ trong nền kinh tế toàn cầu hiện nay.

Résumé

La naissance et le développement vigoureux de la technologie informatique ont permis la naissance et le développement en grand nombre des autres activités scientifiques et technologiques dont les activités de la linguistique. Parallèlement au développement ininterrompu des générations d'ordinateur et des logiciels de traduction automatique, dans le domaine de la linguistique appliquée on voit apparaître une autre tendance de développement : la linguistique de corpus.

La réalité a prouvé que la linguistique de corpus joue un rôle de plus en plus important dans l'économie. Cet article aborde le rôle de la linguistique de corpus, une science qui a vu le jour au milieu de XXème siècle, comme une contribution des sciences sociale et humaine et de la science de la linguistique de corpus au développement actuel de l'économie mondiale.

[1] Hội ngôn ngữ học Việt Nam

1. Cơ sở khái niệm

Sự ra đời và phát triển mạnh mẽ của công nghệ thông tin đã dẫn tới sự ra đời và phát triển của hàng loạt các lĩnh vực hoạt động khoa học và công nghệ khác, trong đó có các lĩnh vực hoạt động của ngôn ngữ học. Song song với sự phát triển không ngừng của các thế hệ công nghệ máy tính và dịch tự động, trong ngôn ngữ học hình thành xu hướng phát triển mới - Ngôn ngữ học khối liệu (Corpus Linguistics).

Ngôn ngữ học khối liệu ngày nay là lĩnh vực khoa học hiện đại và đang phát triển rất nhanh. Ngôn ngữ học khối liệu được hình thành xuất phát từ các nhu cầu ngày càng tăng của khoa học ngôn ngữ trong việc áp dụng công nghệ máy tính vào việc xử lý khối lượng lớn các nguồn ngữ liệu.

Ngôn ngữ học khối liệu là khoa học liên ngành giữa ngôn ngữ học, khoa học máy tính và kỹ thuật số, có mối quan hệ trực tiếp với quá trình xây dựng và hoàn thiện các khối liệu văn bản, cũng như với việc sử dụng chúng như một công cụ trong quá trình nghiên cứu ngôn ngữ.

Từ "khối liệu" lần đầu tiên được sử dụng như một thuật ngữ khoa học vào năm 1961[2] để chỉ khái niệm cơ bản của ngôn ngữ học khối liệu. Về nguyên tắc, một tập hợp bất kì các văn bản đều có thể được gọi là khối liệu. Theo tiếng La tinh, khối liệu có nghĩa là "any body of text"[3] (khối văn bản bất kì - ĐHT dịch). Tuy nhiên, thuật ngữ "khối liệu" khi được sử dụng trong ngữ cảnh cụ thể của ngôn ngữ học hiện đại, sẽ có ý nghĩa đặc trưng hơn nhiều so với định nghĩa đơn giản vừa nêu trên. Nếu nhìn nhận từ góc độ khối liệu là cơ sở của ngôn ngữ học khối liệu - khoa học nghiên cứu các phương pháp xây dựng và sử dụng khối liệu với sự trợ giúp của công nghệ máy tính, - có thể dựa vào bốn đặc điểm cơ bản sau đây để định nghĩa khối liệu:

Bao gồm các model điển hình. Nếu là khối liệu của hai ngôn ngữ thì cần bao gồm các model tương đồng điển hình;

Có kích cỡ xác định;

Ở dạng đọc được trên máy tính;

Có các chú giải chuẩn về mặt ngôn ngữ.

Ở đây, khối các văn bản là một khái niệm quan trọng trong ngôn ngữ học khối liệu. Khối liệu là tập hợp số lượng lớn các văn bản của nhiều tác giả và theo nhiều thể loại khác nhau, bao gồm các câu được gán nhãn cấu trúc cú pháp và từ vựng theo nguyên tắc nhất định.

Có thể nói rằng ngôn ngữ học khối liệu xuất hiện vào đầu thập kỉ 60 của thế kỉ XX cùng với sự xuất hiện khối liệu đầu tiên tại Mỹ và bắt đầu phát triển trong vòng vài thập kỉ gần đây. Căn cứ vào bản chất và hoạt động ngôn ngữ của khối liệu, có thể định nghĩa khối liệu là tập hợp các dữ liệu tương đồng về mặt ngôn ngữ, được trình bày dưới dạng model văn bản điện tử, theo các cấu trúc nhất định và được

[2] Thuật ngữ được sử dụng lần đầu tiên trong Brown corpus năm 1961 với gần 1 triệu từ và cụm từ Anh - Mĩ.

[3] Лингвистический энциклопедический словарь. Главн. ред. В.Н. Ярцева. М., 1990. - 685 с.

sử dụng để giải quyết các vấn đề ngôn ngữ cụ thể. Trong trường hợp cụ thể, khối liệu ngôn ngữ bao gồm cả hệ thống điều chỉnh dữ liệu của văn bản nhằm giúp người sử dụng tìm kiếm được các thông tin cần thiết một cách nhanh chóng và dễ dàng. Đối với các nhà nghiên cứu ngôn ngữ, sử dụng khối liệu sẽ tiết kiệm được rất nhiều thời gian và công sức.

2. Vai trò của khối liệu ngôn ngữ trong nghiên cứu và giảng dạy

Hiện nay, các kiểu khối liệu khác nhau đã và đang được xây dựng cho nhiều ngôn ngữ trên thế giới, **với tầm quan trọng và giá trị sử dụng rất lớn**. Đơn cử ví dụ: một mặt, khối liệu ngôn ngữ đã được gán nhãn là nguồn kiến thức được hệ thống hóa quan trọng về cú pháp và được các nhà ngôn ngữ học sử dụng khi tiến hành các nghiên cứu về ngôn ngữ cơ bản; mặt khác, khối liệu ngôn ngữ đã được gán nhãn chính là nguồn tài nguyên quan trọng nhất đối với khoa học về ngôn ngữ máy tính bởi vì nhờ các khối liệu ngôn ngữ này, có thể xây dựng được các chương trình xử lý ngôn ngữ tự nhiên.

Nghiên cứu cho thấy khối liệu ngôn ngữ có các đặc điểm rất đặc trưng và hữu hiệu, trong đó, khối liệu văn bản

- là thành phần không thể thiếu đối với hệ thống dịch máy;
- là tài nguyên ngôn ngữ chuẩn được hình thành và sử dụng trên máy tính;
- cho phép sử dụng các chương trình tự động xử lý dữ liệu theo các chuẩn nhất định;
- cho phép lựa chọn các dữ liệu cần thiết để nghiên cứu và sử dụng.

Trong lĩnh vực nghiên cứu và giảng dạy ngôn ngữ, trên cơ sở khối liệu, người sử dụng có thể rất nhanh chóng nhận biết được

- tần số sử dụng của từ vị, các phạm trù ngữ pháp;
- sự thay đổi tần số sử dụng của từ và cụm từ;
- sự thay đổi tần số của ngữ cảnh văn bản theo lịch đại và đồng đại;
- cách sử dụng ngôn ngữ của các tác giả khác nhau
- và v.v.

Cho đến nay, ngôn ngữ học khối liệu ngày càng có xu hướng phát triển mạnh mẽ cùng với sự phát triển của công nghệ thông tin. Là một bộ phận của ngôn ngữ học hiện đại, ngôn ngữ học khối liệu hiện nay đang được nâng cao hiệu quả về thực hành và hoàn thiện về lí thuyết. Ngôn ngữ học khối liệu đóng vai trò ngày càng quan trọng trong nền kinh tế toàn cầu khi các lĩnh vực khoa học và công nghệ phát triển mạnh. Ngôn ngữ khối liệu là ngôn ngữ bất kì tham gia vào thành phần khối liệu. Có thể nói rằng khối liệu ngôn ngữ đang được sử dụng rộng rãi bởi các nhà ngôn ngữ ứng dụng, các chuyên gia ngôn ngữ - lí luận, ngôn ngữ máy tính, các giảng viên và các chuyên gia thuộc nhiều lĩnh vực khoa học và đời sống khác nhau.

3. Ngôn ngữ học khối liệu trong nền kinh tế toàn cầu

Hiện nay, khi Việt Nam đã gia nhập WTO và xuất hiện sự cần thiết phải thực hiện giao lưu để trao đổi thông tin ở mức độ giao tiếp bằng các ngôn ngữ trên phạm vi toàn thế giới thì điều thiết yếu trong việc hội nhập kinh tế thế giới là cần có **hệ thống khối liệu quốc gia** nhằm phục vụ các lĩnh vực liên quan đến nghiên cứu khoa học, giảng dạy, cập nhật thông tin trong và ngoài nước v.v. Một vấn đề quan trọng khác là trong các điều kiện hiện nay, khi "người phiên dịch chuyên nghiệp cần phải biết vô vàn các thuật ngữ của nhiều chuyên ngành và cần phải thuộc rất nhiều tên gọi chính xác các chủng loại chi tiết, linh kiện, dụng cụ, cơ cấu, các chất v.v. khác nhau[4]", thì một phiên dịch viên dù giỏi đến đâu cũng không thể cập nhật hết được một lượng thông tin khổng lồ trong nền kinh tế toàn cầu phát triển như vũ bão. Lúc này, **việc sử dụng khối liệu ngôn ngữ để trợ giúp cho quá trình dịch thuật là tất yếu và cần thiết.**

Khối liệu trong nghiên cứu ngôn ngữ là công cụ để xây dựng, điều chỉnh và bổ sung các hệ thống tự động hóa khác nhau như dịch tự động, nhận dạng lời nói, tìm kiếm thông tin. Ví dụ, tìm kiếm trong khối các dữ liệu theo một từ bất kì có thể tạo ra được cả một danh mục liệt kê tất cả các trường hợp có sử dụng từ đó với đầy đủ thông tin về nguồn gốc dữ liệu.

Tại nhiều nước trên thế giới như Anh, Mỹ, Nhật, Đức, Nga, Trung Quốc v.v., vấn đề nghiên cứu và sử dụng hữu hiệu các khối liệu ngôn ngữ (language corpora) đã và đang nhận được sự quan tâm đặc biệt từ phía các cơ quan quốc gia. Chất lượng website của các nước này là ví dụ điển hình. Một ví dụ khác là việc dạy và học tiếng Anh ngày nay đạt hiệu quả, trong đó một phần đáng kể là có sự trợ giúp của công nghệ máy tính với việc sử dụng các khối liệu ngôn ngữ. Có thể kể đến các khối liệu quan trọng như Bank of English 1997 với 320 triệu đơn vị từ và cụm từ sử dụng hoặc ICLE 1997 với 200 triệu đơn vị từ và cụm từ sử dụng dưới dạng viết dành cho người nước ngoài[5].

Trong thập kỉ vừa qua, tại nhiều quốc gia đã và đang tiến hành việc xây dựng các khối liệu ngôn ngữ trên cơ sở bản ngữ. Trong đó, mạnh mẽ hơn cả là công trình xây dựng các khối liệu tiếng Anh, xuất hiện lần đầu tiên vào những năm 60 thế kỉ XX, điển hình là Khối liệu Brown và Khối liệu Lancaster/Oslo-Bergen (LOB). Mỗi khối liệu chứa khoảng 1 triệu đơn vị từ và cụm từ sử dụng với sơ đồ hình thái học. Ngoài ra, Khối liệu Lancaster/Oslo-Bergen còn chứa 2 khối liệu con là Leeds-Lancaster Treebank và Khối liệu Lancaster Parsed với sơ đồ cú pháp học. Khối liệu Anh Quốc (BNC) chứa đến 100 triệu đơn vị từ và cụm từ sử dụng cũng được coi là một trong số khối liệu ngôn ngữ lớn nhất hiện nay. Khối liệu này được xây dựng vào những năm 90 thế kỉ XX trên cơ sở sơ đồ hình thái học, bao gồm khoảng 90% đơn vị từ và cụm từ sử dụng ở dạng viết, 10% số đơn vị còn lại ở dạng nói.

---

[4] *Беляева Л.Н.* Теория и практика перевода. Санкт-Петербург, 2003, с.19.

[5] *Рыков В.В.* Корпус текстов как отражение состояния русского языка // Труды Международного конгресса "Русский язык: исторические судьбы и современность" . – Москва: МГУ, 2001 г.

Ngoài các khối liệu ngôn ngữ kể trên, còn tồn tại hàng loạt các khối liệu tiếng Anh khác được sử dụng cho việc nghiên cứu bằng tiếng Anh, cho việc dạy và học tiếng Anh như một ngoại ngữ.[6]

Đối với các nước châu Âu khác, trong số các khối liệu ngôn ngữ có trữ lượng lớn và giá trị sử dụng cao, cần kể đến Khối liệu tiếng Đức. Đây là tập hợp lớn nhất các văn bản và ngôn bản bằng tiếng Đức, bao gồm khoảng 2 tỉ đơn vị từ và cụm từ sử dụng. Khối liệu này chứa sơ đồ hình thái-cú pháp học dựa trên cơ sở SGML (Standard Generalized Markup Language). Hệ thống tự động hóa COSMAS II của khối liệu tiếng Đức cho phép người sử dụng dễ dàng tìm kiếm thông tin chứa trong khối liệu này theo các dấu hiệu tình thái học của dạng từ. Một hệ thống khác cũng cần kể đến là Khối liệu tiếng Tiệp với 100 triệu đơn vị từ và cụm từ sử dụng. Ở đây, chương trình ngôn ngữ hỗ trợ cho khối liệu là chương trình tạo lập danh mục từ và cụm từ trong khối liệu, cho phép cập nhật toàn bộ các ví dụ sử dụng với đầy đủ trích dẫn, tần số xuất hiện, phân tích ngữ pháp từ hoặc cụm từ sử dụng trong khối liệu.[7]

Đối với các nước châu Á, Trung Quốc và Nhật Bản là những nước có các khối liệu bản ngữ lớn nhất. Khối liệu tiếng Trung chứa khoảng 1 tỷ đơn vị từ và cụm từ, đang được sử dụng rất rộng rãi và hữu hiệu.[8]

Tại Nga, ngôn ngữ học khối liệu được bắt đầu nghiên cứu mới chỉ trong vòng hai thập kỉ trở lại đây, nhưng với tốc độ rất nhanh về thực hành, chuẩn xác về lí thuyết. Hiện nay, khoa học về khối liệu ngôn ngữ đang được giảng dạy tại các trường đại học lớn và nghiên cứu tích cực tại các viện nghiên cứu ngôn ngữ của Liên bang Nga nhằm phục vụ cho một nền kinh tế tăng trưởng. Trong vòng 10 năm trở lại đây, ngôn ngữ học khối liệu được đặc biệt quan tâm nghiên cứu và phát triển. Các khối liệu ngôn ngữ tại Nga được sử dụng rộng rãi trong các lĩnh vực của ngôn ngữ học ứng dụng, từ vựng học, dạy và học ngoại ngữ, ngôn ngữ học máy tính và các lĩnh vực khoa học xã hội khác. Khối liệu tiếng Nga đến nay đã tăng đáng kể về khối lượng các đơn vị từ và cụm từ sử dụng, mở rộng phạm vi sử dụng ngôn ngữ trong nhiều lĩnh vực khoa học khác nhau.

Đối với Việt Nam, việc nghiên cứu và xây dựng Khối liệu tiếng Việt (nội dung cụ thể sẽ được trình bày trong một bài báo khác) là cần thiết và cấp bách trong nền kinh tế hội nhập hiện nay. Khối liệu tiếng Việt có quan hệ trực tiếp đến các hoạt động xã hội, do đó, sẽ đem lại hiệu quả cho các hoạt động nói trên. Khối liệu tiếng Việt sẽ phát huy vai trò và tác dụng trong lĩnh vực quảng cáo các thương hiệu của Việt Nam trên thương trường quốc tế. Nghiên cứu và xây dựng các khối liệu ngôn ngữ đòi hỏi phải xác định và chuyển chính xác nghĩa của từng văn bản cụ thể vào khối liệu nhằm giúp người sử dụng cập nhật chính xác thông tin tìm kiếm.

---

[6] http://www.viniti.ru

[7] *McEnery T., Wilson A.* Corpus Linguistics. – Edinburgh: Edinburgh University Press, 1999.

[8] http://ru.wikipedia.org

Trong điều kiện thông tin quốc tế, sự cần thiết xây dựng các khối liệu song song tiếng Việt - tiếng nước ngoài – tiếng Việt liên quan trực tiếp đến các lĩnh vực dịch thuật và dạy-học ngoại ngữ do các nguyên nhân chủ yếu sau đây:

Số lượng sách đọc bằng tiếng nước ngoài trong các thư viện rất lớn, trong khi số người vào thư viện để ngồi đọc sách là không đáng kể;

Phần lớn học sinh, sinh viên Việt Nam học ngoại ngữ hoặc người nước ngoài học tiếng Việt đều có nhu cầu nắm vững các cấu trúc ngôn ngữ tương đương để có thể giao tiếp được bằng tiếng nước ngoài hoặc tiếng Việt khi cần thiết;

Phần lớn các chuyên gia có nhu cầu đọc nhanh tài liệu là nguyên bản hoặc đã được dịch sang một ngôn ngữ khác (ví dụ, văn bản tiếng Việt và bản dịch sang tiếng Anh);

"Rào ngăn cách" ngôn ngữ còn đang tồn tại trong cộng đồng và làm cản trở việc truy cập thông tin từ các website không có hỗ trợ sử dụng bằng tiếng Việt.

4. Lời kết

Việc Việt Nam gia nhập WTO cũng có nghĩa là tiếng Việt gia nhập "cộng đồng ngôn ngữ" của các dân tộc trên thế giới. Vai trò của ngôn ngữ học khối liệu đã được các nghiên cứu trong nước và quốc tế về khoa học này trong thời đại ngày nay đề cập và làm sáng tỏ. Nghiên cứu, xây dựng và sử dụng các khối liệu cùng ngôn ngữ của nó (ngôn ngữ khối liệu) là một giải pháp để đẩy nhanh tiến độ hội nhập về kinh tế, xã hội và chính trị.

Tài liệu tham khảo

McEnery T., Wilson A. *Corpus Linguistics*. – Edinburgh: Edinburgh University Press, 1999.

Марчук Ю.Н. *Корпус текстов и сверхбольшие базы лингвистических данных* // Сборник: Труды международной конференции «Корпусная лингвистика – 2002». - Издательство Санкт-Петербургского университета, 2002.

Шимкова М. *Репрезентативность корпуса как лингвистическая проблема* // Сборник: Труды международной конференции «Труды международной конференции «MegaLing'2005. Прикладная лингвистика в поиске новых путей» – 2005.

Клименко С.В., Рыков В.В. *Логические индукция и дедукция как принципы отражения предметной области в корпусе текстов* // Труды Международного семинара Диалог ‘2001 по компьютерной лингвистике и ее приложениям. – Аксаково, 2001 г.

Апресян Ю.Д., Иомдин Л.Л., Санников А.В., Сизов В.Г. *Семантическая разметка в глубоко аннотированном корпусе русского языка*. // Сборник: Труды международной конференции «Труды международной конференции «MegaLing'2005. Прикладная лингвистика в поиске новых путей» – 2005.

Баранов А.Н. *Проблема репрезентативности корпуса данных (на примере политической метафорики)* // Труды Международного семинара Диалог ‘2001 по компьютерной лингвистике и ее приложениям. – Аксаково, 2001 г.

Милчонока Э. *Обзор ресурсов латышского языка в Институте математики и информатики Латвийского университета* // Сборник: Труды международной конференции «Корпусная лингвистика – 2002». - Издательство Санкт-Петербургского университета, 2002.

Беляева Л.Н. *Теория и практика перевода*. Санкт-Петербург, 2003.

Жукова В.В. *К вопросу об интенсификации процесса обучения взрослых иностранному языку (на материале английского языка)*. - с. 291 - 303. // Функциональные стили и преподавание иностранных языков. Отв. ред. М.Я. Цвиллинг. М., Наука, 1982. - 360 с.

Лингвистический энциклопедический словарь. Главн. ред. В.Н. Ярцева. М., 1990. - 685 с.

Розенталь М.А., Теленкова М.А. *Словарь – справочник лингвистических терминов*. М., “Просвещение”, 1985. – 399 с.

Дао Хонг Тху. *Корпус параллельных текстов в аспекте корпусной лингвистики //* Проблемы современной филологии и лингводидактики, сб. научных трудов, СПб, изд.РГПУ им.А.И.Герцена, 2006, с.23-28.

Đào Hồng Thu (2009). *Ngôn ngữ học khối liệu và những vấn đề liên quan (Quyển I)*. Nxb. Khoa học xã hội, Hà Nội.

Đào Hồng Thu (2010). *Hướng sử dụng khối liệu tiếng Việt. Báo cáo tại Hội thảo về Nghiên cứu, phát triển các sản phẩm công nghệ xử lý tiếng Việt tháng 8 năm 2010*. Bộ Khoa học và Công nghệ.