TIẾNG VIỆT VÀ CHỮ VIỆT

# Ngôn ngữ học khối liệu (Corpus)

(Phần 1)

**Đào Hồng Thu**
**(TS, Hà Nội)**

**1. Dẫn nhập**

Thực tế đã chứng minh rằng khoa học về ngôn ngữ luôn gắn liền với các thành tựu của khoa học kĩ thuật và công nghệ. Sự ra đời và phát triển của máy tính đã dẫn đến sự hình thành và phát triển của nhiều lĩnh vực khoa học, trong đó có các lĩnh vực của ngôn ngữ học ứng dụng.

Trong những năm nửa cuối thế kỉ XX và đầu thế kỉ XXI, cùng với sự phát triển của khoa học thông tin, khoa học ngôn ngữ liên tục phát triển và hình thành các xu hướng phát triển mới nhằm đáp ứng nhu cầu hoạt động của xã hội. Song song với sự phát triển không ngừng của các thế hệ công nghệ máy tính và dịch tự động, trong ngôn ngữ học ứng dụng hình thành xu hướng phát triển mới - **Ngôn ngữ học Corpus (Ngôn ngữ học khối liệu)**.

Ngôn ngữ học Corpus (Ngôn ngữ học khối liệu) là ngành khoa học trẻ, là giao điểm giữa khoa học ngôn ngữ và khoa học máy tính, được hình thành vào cuối thế kỉ XX trên cơ sở kĩ thuật điện tử 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.

Có thể dẫn chứng một ví dụ đơn giản về vai trò và sức sống của ngành khoa học này. Bất kì nhà ngôn ngữ nào khi nghiên cứu cũng gặp phải vấn đề về việc lựa chọn ngữ liệu cho đề tài nghiên cứu của mình, nghĩa là cần tham khảo rất nhiều loại văn bản để tìm ra các ví dụ cần thiết, và phải hài lòng với việc ngẫu nhiên lựa chọn được các ví dụ đó. May mắn là hiện nay đã có nhiều văn bản bằng các ngôn ngữ khác nhau có thể tìm kiếm được ở dạng văn bản điện tử (file của máy tính). Khả năng sử dụng các nguồn ngữ liệu trên làm dễ dàng rất nhiều quá trình tìm kiếm thô sơ, đồng thời đòi hỏi nâng cao hơn chất lượng nghiên cứu, nghĩa là số lượng ví dụ tìm kiếm được cần đầy đủ hơn nhiều cho mỗi ngôn ngữ được nghiên cứu. Tuy nhiên, làm việc với các văn bản trên file máy tính cũng không kém nhọc nhằn. Để có thể khắc phục sự mệt nhọc không cần thiết trong công việc của nhà nghiên cứu, khắp nơi trên thế giới đã thành lập các chương trình về khối liệu. Các chương trình đặc thù này có thể đáp ứng rất nhiều yêu cầu của người sử dụng, ví dụ, một chương trình về văn học Việt Nam thế kỉ XX có thể đưa ra toàn bộ các câu, tập hợp câu hoặc văn bản có chứa tập hợp từ "*văn học Việt Nam*" được đăng trên các báo, tạp chí v.v. Nghiên cứu và tạo lập các chương trình khối liệu như trên là nhiệm vụ của Ngôn ngữ học khối liệu.

**2. Các khái niệm cơ bản**

Từ "corpus" (với nghĩa là "khối liệu") lần đầu tiên được

sử dụng như một thuật ngữ khoa học vào năm 1961[1] để chỉ khái niệm cơ bản của Ngôn ngữ học khối liệu. Thuật ngữ này được dùng trong tập hợp các văn bản bằng các ngôn ngữ khác nhau dưới dạng văn bản điện tử (file của máy tính): khối liệu Brown, khối liệu London-Lund v.v.

Các nhà nghiên cứu người Anh T. McEnery và A. Wilson đã đưa ra định nghĩa chung cho khái niệm *khối liệu* như sau:

a. (*sử dụng tự do*) khối liệu là văn bản bất kì;

b. (*sử dụng thường xuyên*) khối liệu là văn bản điện tử;

c. (*sử dụng theo phong cách ngôn ngữ*) khối liệu là văn bản điện tử, được tập hợp sao cho có sự hiện diện của tất cả các phong cách ngôn ngữ chức năng.[2]

Có thể coi một tập hợp bất kì các văn bản là khối liệu. Theo tiếng La tin, 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, cụ thể là trong ngôn ngữ học máy tính, 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, - thì 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.

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 sử dụng để giải quyết các vấn đề ngôn ngữ cụ thể. Khối liệu trong ngôn ngữ học máy tính 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.

Khối liệu 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 liệu 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. Đối với các nhà nghiên cứu ngôn ngữ, sử dụng khối liệu sẽ tiết kiệm được rất nhiều thời gian và công sức.

Khối liệu văn bản là cần thiết và hữu ích đối với giới ngôn ngữ học hiện đại bởi vì chúng tạo ra những khả năng mới cho việc nghiên cứu của các nhà ngôn ngữ, làm tiết kiệm đáng kể thời gian và đảm bảo cập nhật được lượng lớn thông tin một cách rất nhanh chóng. Nhờ khối liệu có thể trong vài giây biết

---

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

[2] *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. – С.97.

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

được tần số sử dụng của các loại từ và cụm từ cần nghiên cứu, theo dõi thường xuyên và điều chỉnh được tần số xuất hiện của chúng trên các phương tiện thông tin khoa học và đại chúng.

Tìm kiếm dữ liệu trong khối liệu cho phép trên cơ sở một từ bất kì tạo ra được danh mục của tất cả các trường hợp sử dụng của từ đó trong ngữ cảnh với nguồn trích dẫn đầy đủ. Các khối liệu có thể được sử dụng để nhận biết các thông tin hướng dẫn, tham khảo và số liệu thống kê về các đơn vị ngôn ngữ và lời nói. Khối liệu có thể cung cấp cho người sử dụng các thông tin về tần số hoạt động của từ và cụm từ, lexeme và v.v.

Khối liệu cho phép theo dõi các thay đổi về tần số sử dụng các đơn vị từ vựng và các ngữ cảnh ở các giai đoạn phát triển khác nhau của lịch sử xã hội loài người. Khi nhận được các dữ liệu ngôn ngữ trong một giai đoạn phát triển lịch sử nhất định từ khối liệu, người sử dụng có thể nghiên cứu các quá trình biến đổi thành phần từ vựng của ngôn ngữ trên thực tế, có thể tiến hành các phân tích cú pháp ở các thể loại văn bản và của các tác giả khác nhau.

Khối liệu còn được sử dụng làm cơ sở cho việc chuẩn bị các loại từ điển hiện đại và lịch sử khác nhau một cách nhanh chóng và hiệu quả. Vai trò của Ngôn ngữ học khối liệu càng được khẳng định khi các công trình nghiên cứu về khối liệu cho thấy khối liệu có thể sử dụng để xây dựng các kĩ năng và kiểm tra ngữ pháp trong quá trình dạy học ngoại ngữ và dịch thuật.

**3. Lược sử quá trình hình thành và phát triển của Ngôn ngữ học khối liệu**

Xuất phát điểm của sự hình thành và ra đời Ngôn ngữ học khối liệu có thể tính vào thời điểm đầu những năm 60 thế kỉ XX, khi xuất hiện khối liệu văn bản ngôn ngữ đầu tiên tại Mĩ và bắt đầu phát triển trong vòng hai thập kỉ trở lại đây. Năm 1963, lần đầu tiên khối liệu văn bản điện tử - khối liệu Brown được xây dựng tại trường đại học Brown (Mĩ) do các tác giả là W. Francis và H. Kucera thiết kế và xây dựng bao gồm 1 triệu đơn vị từ và cụm từ Anh - Mĩ từ các văn bản in ấn được lựa chọn vào năm 1961. Sự xuất hiện của khối liệu Brown đã gây sự quan tâm lớn không những đối với các nhà ngôn ngữ học, trước hết, về các nguyên tắc lựa chọn văn bản và các nhiệm vụ được giải quyết trong khối liệu.

Tiếp theo khối liệu Brown là sự ra đời của hàng loạt các khối liệu. Các nghiên cứu cho thấy rằng Ngôn ngữ học khối liệu được hình thành như một ngành khoa học độc lập về ngôn ngữ văn bản là vào những năm 90 thế kỉ XX. Ngôn ngữ học khối liệu vẫn có các mối quan hệ mật thiết với Ngôn ngữ học máy tính qua việc sử dụng các thành tựu của Ngôn ngữ học máy tính và ngược lại, gây ảnh hưởng tích cực lên Ngôn ngữ học máy tính trong quá trình phát triển.

Trong thập kỉ vừa qua, tại nhiều quốc gia đã và đang tiến hành việc xây dựng các khối liệu trên cơ sở bản ngữ. Trong đó, mạnh mẽ hơn cả là công trình xây dựng khối liệu tiếng Anh, xuất hiện lần đầu tiên vào những năm 60 thế kỉ XX, điển hình sau khối liệu Brown University là khối liệu Lancaster/Oslo-Bergen khối liệu (LOB). Mỗi khối liệu chứa khoảng 1 triệu đơn vị từ và cụm từ sử dụng với sơ đồ hình thái học. Ngoài ra, Lancaster/Oslo-Bergen khối liệu còn chứa 2 khối liệu con là các khối liệu Leeds-Lancaster Treebank và Lancaster Parsed khối liệu 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 các khối liệu 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.

Ngày nay, việc dạy và học tiếng Anh 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. 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[4]. Ngoài các khối liệu kể trên, còn tồn tại hàng loạt 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.[5]

Đối với các nước châu Âu khác, trong số các khối liệu, 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 thống 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.[6]

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, phục vụ đắc lực cho nền kinh tế phát triển của Trung Quốc.[7]

Tại Liên bang Nga, ngôn ngữ học khối liệu được bắt đầu nghiên cứu mới chỉ trong vòng hơn thập kỉ trở lại đây, nhưng với tốc độ rất nhanh về thực hành, chuẩn xác về lí thuyết. Hiện nay, ngôn ngữ học khối liệu 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 5-6 năm trở lại đây, Ngôn ngữ học ở LB Nga khối liệu được đặc biệt quan tâm nghiên cứu và phát triển. Khối liệu tại LB 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 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.

Ở Việt Nam, việc xây dựng khối liệu tiếng Việt trong

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

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

[6] *McEnery T., Wilson A.* Khối liệu Linguistics. – Edinburgh: Edinburgh University Press, 1999.

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

quá trình hội nhập quốc tế của Việt Nam là vấn đề cần thiết và cấp bách.

Nhờ sự phát triển của các khối văn bản tương đương giữa các cặp ngôn ngữ, cuối thế kỉ XX đã xuất hiện hệ thống dịch theo phương pháp thống kê tự động đầu tiên, «…mặc dù vẫn còn những hạn chế, phương pháp thống kê đối với việc dịch tự động đã làm giảm nhẹ đáng kể so với việc xây dựng các hệ thống theo phương pháp truyền thống. Thành tựu không thể phủ nhận của các hệ thống này là loại bỏ việc xây dựng các từ điển điện tử theo phương pháp thủ công …»[8].

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 ứng dụng, 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. Có thể nói rằng khối liệu 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.

[8] Марчук Ю.Н. Корпус текстов и сверхбольшие базы лингвистических данных // Сборник: Труды международной конференции «Корпусная лингвистика – 2002». - Издательство Санкт-Петербургского университета, 2002. – С.96.

**Tài liệu tham khảo**


1. *Brown, R.* (1973) A First Language: The Early Stages, Cambridge, MA: Harvard University Press.
2. *Chomsky, N.* (1968) Language and Mind, Harcourt Brace, New York.
3. *Mcenery, T. and Wilson, A.* (1996) Khối liệu Linguistics. Edinburgh University Press.
4. *Barnbrook, G.* (1996). Language and Computers: a practical introduction to the computer analysis of language. Edinburgh University Press.
5. *Woods, A., Fletcher, P., and Hughes, A.* (1986). Statistics in Language Studies. Cambridge. Cambridge University Press.
6. *McEnery T., Wilson A.* (1999). Khối liệu Linguistics. – Edinburgh: Edinburgh University Press.
7. *Марчук Ю.Н.* (2002). Корпус текстов и сверхбольшие базы лингвистических данных // Сборник: Труды международной конференции «Корпусная лингвистика – 2002». - Издательство Санкт-Петербургского университета.
8. *Милчонока Э.* (2002). Обзор ресурсов латышского языка в Институте математики и информатики Латвийского университета// Сборник: Труды международной конференции «Корпусная лингвистика – 2002». - Издательство Санкт-Петербургского университета.
9. *Рыков В.В.* (2001). Корпус текстов как отражение состояния русского языка // Труды Международного конгресса "Русский язык: исторические судьбы и современность" . – Москва: МГУ.
10. Лингвистический энциклопедический словарь. Главн. ред. В.Н. Ярцева. М., 1990. - 685 с.
11. *Розенталь М.А., Теленкова М.А.* (1985). Словарь – справочник лингвистических терминов. М., “Просвещение”. – 399 с.
12. *Дао Хонг Тху* (2006). Корпус параллельных текстов в аспекте корпусной лингвистики. // Проблемы современной филологии и лингводидактики, сб. научных трудов, СПб, изд.РГПУ им. А.И.Герцена, с.23-28.


(Bài này gửi đến Ban biên tập ngày 06-06-2007)