NGÔN NGỮ HỌC

KHỐI LIỆU TRONG XU THẾ TOÀN CẦU HÓA

HIỆN NAY, KHI CHÚNG TA 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 VIỆT NAM 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... VỚI TỐC ĐỘ PHÁT TRIỂN CỦA THÔNG TIN KHOA HỌC VÀ CÔNG NGHỆ NHƯ HIỆN NAY, 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.

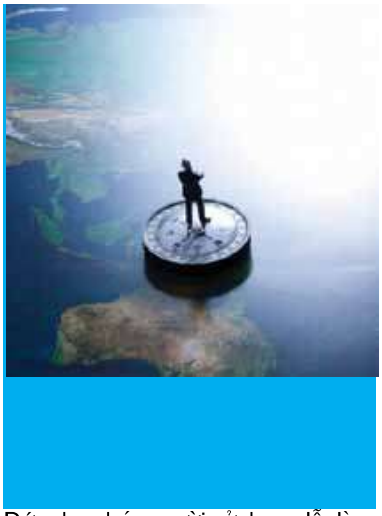
Ngôn ngữ học khối liệu xuất hiện vào đầu thập kỉ 60 của thế kỉ XX cùng với sự xuất hiện khối liệu đầu tiên tại Mĩ và bắt đầu phát triển trong vòng hai thập kỉ trở lại đây. Cho đến nay, ngôn ngữ học khối liệu ngày càng có xu hướng phát triển mạnh mẽ cùng với sự phát triển của công nghệ thông tin. Là một 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. 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.

Nói đến khối liệu là nói đến 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. Tại nhiều nước

trên thế giới như Anh, Mĩ, Nhật, Đức, Nga, Trung Quốc v.v., vấn đề nghiên cứu và sử dụng hữu hiệu các khối liệu (corpora) đã và đang nhận được sự quan tâm đặc biệt từ phía các quốc gia. Chất lượng website 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 nhờ sự trợ giúp của công nghệ máy tính với việc sử dụng corpora. 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.

Trong thập kỉ vừa qua, tại nhiều quốc gia đã và đang tiến hành việc xây dựng corpora trên cơ sở bản ngữ. Trong đó, mạnh mẽ hơn cả là công trình xây dựng corpora tiếng Anh, xuất hiện lần đầu tiên vào những năm 60 thế kỉ XX, điển hình là Khối liệu Brown University 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 corpora 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. Ngoài corpora kể trên, còn tồn tại hàng loạt corpora tiếng Anh khác được sử dụng cho việc nghiên cứu bằng tiếng Anh, cho việc dạy và học tiếng Anh như một ngoại ngữ.

Đối với các nước châu Âu khác, trong số corpora, 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.

Đối với các nước châu Á, Trung Quốc và Nhật Bản là những nước có corpora bản ngữ lớn nhất. Khối liệu tiếng Trung chứa 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.

Tại 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, khoa học về 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 khối liệu được đặc biệt quan tâm nghiên cứu và phát triển. Corpora 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 lượng đáng kể 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.

Trong điều kiện thông tin quốc tế, sự cần thiết xây dựng corpora tiếng Việt - tiếng nước ngoài 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 hoặc người nước ngoài học tiếng Việt có nhu cầu cần 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 dưới dạng 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 cản trở việc truy cập thông tin từ các website không sử dụng tiếng Việt.

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.

TS. Đào Hồng Thu