NGOẠI NGỮ VỚI BẢN NGỮ

NGHIÊN CỨU KHẢO SÁT BNC VÀ ICE THEO HƯỚNG KHỐI LIỆU TIẾNG ANH VÀ ỨNG DỤNG THỰC TIỄN

SURVEY ON ACTUAL USE OF BNC AND ICE UNDER ENGLISH CORPUS ORIENTATION AND PRACTICAL APPLICATIONS

NGUYỄN ĐÌNH TRƯƠNG NGUYỄN
(Đại học Khoa học, ĐH Huế)

Abstract

British National Corpus (BNC) and the International Corpus of English (ICE) are popular English corpora in language teaching and researching now in many countries in the World. Based on written and spoken texts of the British National Corpus (BNC) and the International Corpus of English (ICE), the author surveys researches on these corpora in order to contribute some ways of practical data analysis and data surveying to applied linguistics.

1. Đặt vấn đề

Những năm gần đây, việc nghiên cứu ngôn ngữ học theo hướng khối liệu là hoàn toàn mới đối với ngôn ngữ học Việt Nam. Tuy vậy, trên thế giới, nghiên cứu lĩnh vực này phát triển rất mạnh, ngôn ngữ học khối liệu (Corpus Linguistics) - NNHKL là khoa học về sáng tạo và phân tích khối liệu ngôn ngữ một cách chặt chẽ.

Việc nghiên cứu này dựa trên lí thuyết ngôn ngữ học, liên hệ đến hệ hình, ngôn ngữ xã hội học, ngôn ngữ tâm lí học. Việc nghiên cứu này giúp chúng ta có cách nhìn mới trong phân tích dữ liệu và so sánh đối chiếu ngữ liệu trong ngôn ngữ, xử lí ngôn ngữ theo văn bản và ngôn bản, góp phần quan trọng vào quá trình dạy - học và nghiên cứu ngôn ngữ, đặc biệt là tiếng Anh được tốt hơn.

2. Phương pháp nghiên cứu

Nghiên cứu này sử dụng số liệu trên cơ sở ngôn ngữ học thống kê, đặc biệt, sử dụng các chuyên ngành liên quan như ngôn ngữ xã hội học (Sociolinguistics) và ngôn ngữ tâm lí học (Psycholinguistics). Trên cơ sở the British National Corpus (BNC) và International Corpus of English (ICE), tác giả bài báo chủ yếu phân tích và chọn lọc, phân tích cấu trúc dữ liệu ngôn bản và văn bản. Bài báo đề cập đến vấn đề khảo sát một số tư liệu và nghiên cứu ngôn ngữ trên văn bản (text) tiếng Anh của BNC, ICE để đánh giá, xếp loại và phân tích.

3. Kết quả nghiên cứu

3.1. BNC-the British National Corpus

Xấp xỉ 100 triệu từ, BNC là một trong số khối liệu lớn nhất, bao gồm khoảng 90% đơn vị từ và cụm từ liên quan đến văn bản và khoảng 10% còn lại liên quan đến ngôn bản ở các thể loại khác nhau.

Bảng kết hợp của BNC (theo Aston và Burnard 1998:29-33 và http://info.ox.ac.uk/bnc/what/balance/html).

Speech type (văn nói)

Thể loại	*số ngôn bản nghiên cứu*	*số từ*	*% khối liệu*	*ngôn bản*
Nhân khẩu (demographically sample)	153		4,211,216	41%
giáo dục	144		1,265,318	12%
thương mại	136		1,321,844	13%
cơ quan	241		1,345,694	13%
tâm sự (leisure)	187		1,459,419	14%
không xác định (unclassified)	54		761,973	7%
Tổng cộng	915		10,365,464	100%

Writing type (dạng văn bản)

Thể loại	*số văn bản nghiên cứu*	*số từ*	*% khối liệu văn bản*
Tưởng tượng	625	19,664,309	22%
Khoa học tự nhiên	144	3,752,659	4%
Khoa học đời sống	364	7,369,290	8%
Khoa học xã hội	510	13,290,411	15%
Lĩnh vực thế giới	453	16,507,399	18%
Thương mại	284	7,118,321	8%
Nghệ thuật	259	7,253,846	8%
Lí thuyết và nhận thức	146	3,053,672	3%
Tâm sự	374	9,990,080	11%
Không xác định	50	1,740,527	2%
Tổng cộng	3,209	89,740,544	99%

Qua bảng số liệu trên,có thể thấy rằng thể loại khác nhau về ngôn bản và văn bản được tập hợp thành một khối liệu ngôn ngữ (khối liệu tiếng Anh). Mỗi thể loại văn bản được tập hợp thành khối liệu con ước tính bao gồm khoảng 40.000 từ. Ví dụ, khối liệu con thể hiện tuổi tác và giới tính được tập hợp chủ yếu từ sách báo, ý kiến của công chúng. Về khối liệu ngôn bản, cơ sở dữ liệu được tập hợp từ các nguồn theo phương ngữ của Anh quốc và thuộc các tầng lớp xã hội khác nhau. Nhìn vào bảng mà theo như Biber (1993:256) nhận định tiến trình như của một chu kì-cyclical process. Theo bảng trên, đòi hỏi có sự cân bằng khối liệu về tuổi tác, giới, tầng lớp xã hội và nguồn gốc, sự phân tích các con số quyết định các dữ liệu cần quan tâm.

Nhìn chung,dung lượng của một khối liệu có vai trò quan trọng trong so sánh nguồn và góp phần xác định thời gian nghiên cứu. Để so sánh, theo nghiên cứu ngôn ngữ học thống kê của chúng tôi, 300 mẫu văn nói và 200 mẫu văn viết đòi hỏi hoàn thành thời gian nghiên cứu là 40 giờ tuần và hơn 3 năm để hoàn thành.

3.2 Các thể loại nghiên cứu (International corpus of English)

Trong nghiên cứu này,ngôn bản chiếm 41% là thuộc mẫu nhân sinh (demographically sampled), bao gồm cả hội thoại và độc thoại, 22% văn bản thuộc tưởng tượng, tiểu thuyết và văn chương cảm hứng sáng tác.Nhìn bảng trên cho thấy tỷ lệ phần trăm khoa học xã hội chiếm 15% trong văn bản, 18% - thuộc lĩnh vực thế giới, nghệ thuật chiếm 8% và khoa học tự nhiên chiếm 4%. Ở bảng sau của ICE (International Corpus of English) đưa ra 2 khối liệu con cùng thể loại, ở đây thể loại có nhiều nét đáng chú ý hơn của BNC,có 60% ngôn bản là hội thoại và 40% là độc thoại.

Bảng ICE (từ Greenbaum 1996a:29-30)-Speech type

Văn nói *Loại*	*số văn bản*	*độ dài*	*% ngữ liệu nói*

*Hội thoại*180	360,000	59%	
Cá nhân	100	200,000	33%
Đàm thoại trực tiếp	90	180,000	30%
Đàm thoại có khoảng cách	10	20,000	3%
Công chúng	80	160,000	26%
Lớp học	20	40,000	7%
Đàm thoại truyền hình	20	40,000	7%
Đàm thoại phỏng vấn	10	20,000	3%
Tranh luận nghị trường	10	20,000	3%
Đối chất trong luật	10	20,000	3%
Công việc thương mại	10	20,000	3%
Độc thoại	120	240,000	40%
Không phải chữ viết	70	140,000	23%
Bài bình luận tự do	20	40,000	7%
Bài nói	30	60,000	10%
Thuyết minh	10	20,000	3%
Trình bày (luật)	10	20,000	3%
Chữ viết (đọc văn bản)	50	100,000	17%
Bản tin truyền hình	20	40,000	7%
Hội thoại truyền hình	20	40,000	7%
Bài nói (không thuộc truyền hình)	10	20,000	3%
Tổng cộng	300	600,000	99%

Writing (văn bản) từ Bảng ICE (từ Greenbaum 1996a:29-30)

Loại	*số văn bản*	*độ dài*	*% ngữ liệu viết*
Không xuất bản	50	100,000	26%
Bài văn (không giới hạn của sinh viên)	10	20,000	5%
luận văn (bài thi của sinh viên)	10	20,000	5%
văn chương (xã hội)	15	30,000	8%
văn bản (kinh tế)	15	30,000	8%
xuất bản	150	300,000	75%
thông tin (nghiên cứu)	40	80,000	20%
khoa học xã hội nhân văn khoa học tự nhiên, công nghệ thông tin (đại chúng)	40	80,000	20%
khoa học xã hội nhân văn khoa học tự nhiên, công nghệ thông tin (phóng sự)	20	40,000	10%
tài liệu: hành chính sự nghiệp, kĩ năng, sở thích	20	40,000	10%
báo xã luận	10	20,000	5%
sáng tạo (tiểu thuyết, truyện)	20	40,000	10%
Tổng cộng	200	400,000	101%

Nghiên cứu trên cho thấy thể loại thương mại, tâm sự (tự sự) được nhấn mạnh trong ngôn bản nhiều hơn. Đa số văn bản của BNC thuộc ngôn bản, còn khối liệu ICE chỉ 40% là văn bản. ICE bao gồm 33% ngôn bản tồn tại trong hội thoại trực tiếp hay cuộc đàm thoại điện thoại, còn BNC chứa 41% ngôn bản thuộc loại này. Tại sao có sự khác biệt về dung lượng ngữ liệu trong khối liệu ICE và BNC? Nhìn chung, hai khối liệu này đều cùng thể loại ngôn bản và văn bản. Câu hỏi đặt ra là tại sao có thể loại này và không có thể loại khác? Câu hỏi này cân nhắc hai loại thể loại chính, vì có thể loại text sinh viên được đưa vào nghiên cứu.Nhìn chung, phụ thuộc vào độ dài của văn bản cá nhân, số liệu của văn bản hay ngôn bản

cần nghiên cứu, sắp xếp trong khối liệu để tiến hành nghiên cứu. Việc sử dụng mẫu phương pháp lựa chọn ngôn bản và văn bản là tùy xác suất nghiên cứu tính toán.

1. Khối liệu lớn là tốt hơn khối liệu nhỏ về độ dài trong nghiên cứu. Tuy nhiên, bảng khối liệu trọng yếu hơn là độ lớn trong xếp loại khối liệu.

2. Xếp loại khối liệu bao gồm khối liệu nhiều vấn đề (a multi-purpose corpus) và khối liệu vấn đề đặc biệt (a special-purpose corpus).

3. Để thực hiện, những khúc đoạn văn bản trong khối liệu là quan trọng hơn văn bản hoàn hảo. Những khúc đoạn có thể ngắn khoảng 2.000 từ, đặc biệt là trọng điểm nghiên cứu thường liên quan đến cấu trúc ngữ pháp.

4. Đa số thể loại cấu trúc cần chắc chắn trên căn cứ nghiên cứu, xác định tốt nhất có bao nhiêu biến đổi nội bộ ở thể loại: nhiều thể loại, nhiều cấu trúc cần thu thập.

5. Xác suất cấu trúc văn bản kĩ thuật có thể được sử dụng để xác định thể loại cấu trúc cần thiết bao gộp trong một khối liệu. Tuy nhiên, kĩ thuật lựa chọn trên những nghiên cứu riêng lẻ được chọn dùng điều tra là tốt hơn.

6. Thật sự không thể lựa chọn những riêng lẻ ở xác suất cấu trúc văn bản kĩ thuật, và cấu trúc đó không có giá trị xác suất có thể được sử dụng, điều đó đa số những biến đổi được cân nhắc và cũng cần được kiểm soát khi nghiên cứu.

3.3. Một vài kiểm soát biến đổi thuộc lĩnh vực xã hội học trong nghiên cứu khối liệu

Sự biến đổi phụ thuộc về những biến đổi thuộc ngôn ngữ xã hội học mà cần được cân nhắc trước khi lựa chọn những mẫu của người nói và viết mà ngôn bản của họ được đưa vào khối liệu cần nghiên cứu. Cần chú ý đến giới, tuổi và trình độ học vấn, phương ngữ là yếu tố cơ bản chủ yếu trong nghiên cứu khối liệu thể loại này.

Cần cân bằng thể loại trong nghiên cứu khối liệu, tuổi tác, trình độ giáo dục, sự khác nhau về phương ngữ, ngữ cảnh xã hội (social context) và mối quan hệ xã hội. Mục tiêu nghiên cứu khối liệu về những địa hạt và xã hội khác nhau. Đó là ý nghĩa mục đích khối liệu thuộc về phương ngữ thuộc địa hạt,vùng đặc biệt và phương ngữ thuộc xã hội khác nhau hơn là sự rộng lớn của nó, như British English và American English.

4. Kết luận

Phân tích khối liệu luôn là điểm trọng yếu, giúp chúng ta nâng cao nghiên cứu ngôn ngữ con người nói chung và nghiên cứu ngôn ngữ học nói riêng. Qua bảng số liệu, phân tích đánh giá khối liệu trên cơ sở lí thuyết của ngôn ngữ học ứng dụng giúp có khái niệm về nghiên cứu phân tích ngôn ngữ xử lí số liệu trong một khối liệu. Mục tiêu nghiên cứu nhằm đáp ứng cơ sở nghiên cứu trong nội hạt này, góp phần đẩy mạnh nghiên cứu nhập môn ngôn ngữ học khối liệu còn đang mới mẻ, còn đang trên đà phát triển ở nước ta và ứng dụng vào phác thảo xây dựng khối liệu (planning the construction of a corpus) trong nghiên cứu khoa học văn bản và ngôn bản.

Tài liệu tham khảo

1. A. Hughes (1986), *Statistics in language studies.* Cambrige University Press.

2. A.Wilson and T.Mc Enery (1994), *Teaching and language corpora. Technical report.* Department of modern English language and linguistics, University of lancaster.

3. B. Douglas(1988), *Variation across speech and writing.* New York: Cambrige University Press.

4. B.Michael. *Internatinal journal of corpus linguistics 4,* (1999) 319-327

5. Đào Hồng Thu (2009), *Ngôn ngữ học khối liệu và những vấn đề liên quan* (Quyển 1). Nxb. KHXH.

6. Nguyễn Đức Dân- Đặng Thái Minh (2000), *Thống kê ngôn ngữ học một số ứng dụng.* NXB GD .

7. S.John. *Corpus, Concordance, collocation*. Oxford University Press (1991).

8. http://www.natcorp.ox.ac.uk/

9. http://ice-corpora.net/ice/index.htm

(Ban Biên tập nhận bài ngày 25-09-2012)